An integrated decision-making framework for sustainable data center operation through intelligent load scheduling

Nuoa Lei, Zhu Cheng*, Zhi Cao, Eric Masanet

1 McCormick School of Engineering and Applied Science, Northwestern University, Evanston, IL, USA
2 College of Architecture and Environment, Sichuan University, Chengdu, China
3 Energy and Materials in Infrastructure and Buildings (EMIB), University of Antwerp, Antwerp, Belgium
4 Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA

*Email: nuoalei@u.northwestern.edu, scchengzhu@126.com, emasanet@ucsb.edu

Abstract. Intelligent load scheduling is an emerging approach that has the potential to facilitate extreme sustainable data center (DC) operation. However, scarcity of straightforward tools in the public domain challenges decision makers performing quantitative analysis of the DC load planning and its potential benefits. In this work, a novel integrated decision-making framework was developed to address this issue, which provides the basis for the multi-objective optimization of carbon-, water-, and economic-intelligent load scheduling. The proposed framework was demonstrated with a case study DC in California, which showed the usefulness of the proposed framework in informing sustainable DC operations.

1. Introduction

Unprecedented growth in global internet traffic is driving the demand for data center (DC) services [1]. Because DCs consume enormous amount of electricity and water to support these data services, the industry is now facing challenges to achieve the aggressive environmental sustainability goals set in face of the global warming and water scarcity [2]. Improving DCs’ energy efficiencies has been a long-time pursuit for operators to reduce their electric bill and environmental impacts [3]. With the growing concern about freshwater availability, water-efficient cooling technologies have become distinguished in locations where DCs face water-constraints [4]. Moreover, computing load shifting was recently trialed by Google in response to the power grid’s carbon intensity [5], which presents new opportunities for DCs to contribute to decarbonization. Looking forward, urgent environmental problems require bold actions from the industry, particularly DC co-optimization for carbon-, water-, and economic-intelligent computing. However, straightforward tools that could incorporate all the factors and the potential benefits of the co-optimization has not been found in the literature, which prevents strategic decision makings for sustainable DC operation.

In this work, an integrated framework was developed to fill the aforementioned research gaps, which is a multi-objective optimization model that could inform the DC load planning based on real-time arrival of computing tasks including urgent nonflexible computing workloads, and flexible workloads that can be shifted within contract time limits (CTLs). The proposed optimization
2. Methodology

2.1. Multi-objective load scheduling

The basic idea of carbon-, water-, and economic-intelligent load scheduling is to shift flexible DC workloads \( L_{t}^{f,T} \), ssj_ops/hour at a given timestamp \( T \) to certain allowable time frame (\( t \in \{T, T + k\} \)), resulting in shifted workloads \( \overrightarrow{\vec{w}} = \{L_{t=T}^{f,T}, L_{t=T+1}^{f,T}, \ldots, L_{t=T+k}^{f,T}\} \) that minimize the DC’s carbon emissions (CEs), water footprint (WF), and operating expenses (OPEX). Once the computing tasks were scheduled, the CEs (kgCO\(_2\)e), and the overall WF (liters) associated with the shifted workloads can be quantified as equation (1) and (2):

\[
f_{CO2}(L_{t}^{f,T}) = \sum_{t \in \{T, T+k\}} \frac{L_{t}^{f,T}}{3.6 \times 10^6 g(E(u_t))} (PUE_t \cdot \lambda_{CO2})
\]

\[
f_{WATER}(L_{t}^{f,T}) = \sum_{t \in \{T, T+k\}} \frac{L_{t}^{f,T}}{3.6 \times 10^6 g(E(u_t))} (PUE_t \cdot \lambda_{WCF} + WUE_t)
\]

where \( PUE_t \) (kWh/kWh) and \( WUE_t \) (liters/kWh) respectively stand for the DC’s PUE and WUE, which can be predicted based on forecasted weather at time \( t \) [3]. \( \lambda_{CO2} \) (kgCO2e/kWh) and \( \lambda_{WCF} \) (liters/kWh) respectively refer to the electricity carbon intensity and water consumption factor (WCF), which can be estimated using historical electricity mix data. \( g(\cdot) \) represents the weighted average performance to power ratio (ssj_ops/W) of servers within the DC, which was modeled as a function of the server workload level \( u_t \) using gaussian process regression, and \( E(u_t) \) is the expectation of the server workload level at time \( t \) based on historical DC workload profile. Note that since the arrival of the DC workloads at a future time \( t \) cannot be predicted accurately, we chose not to formulate \( L_{t}^{f,T} \) into \( g(\cdot) \) here. Indeed, a large portion of uncertainty will be contributed by the nonflexible workloads at time \( t \) either or not \( L_{t}^{f,T} \) was formulated into \( g(\cdot) \), and our formulation won’t seriously affect the optimization result because \( g(\cdot) \) is generally a monotonic function.

Similarly, the DC’s OPEX in US dollars (USDs) incorporating the electricity cost \( R_{t}^{e} \), USD/kWh and the water cost \( R_{t}^{w} \), USD/liter can be derived, as expressed by equation (3):

\[
f_{OPEX}(L_{t}^{f,T}) = \sum_{t \in \{T, T+k\}} \frac{L_{t}^{f,T}}{3.6 \times 10^6 g(E(u_t))} (PUE_t \cdot R_{t}^{e} + WUE_t \cdot R_{t}^{w})
\]

At any timestamp \( T \), a multi-objective optimization problem that takes \( \overrightarrow{\vec{w}} \) as the decision vector can thus be formulated as follows:

\[
\text{minimize } \alpha f_{CO2}(L_{t}^{f,T}) + \beta f_{WATER}(L_{t}^{f,T}) + \gamma f_{OPEX}(L_{t}^{f,T})
\]

subject to: \( \sum_{t \in \{T, T+k\}} L_{t}^{f,T} = L^{f,T} \)

and \( 0 \leq L_{t}^{f,T} \leq \Phi L^{DC} - \sum_{P < T} L_{t}^{f,P} \)

where \( \alpha, \beta, \gamma \geq 0 \), and \( \alpha + \beta + \gamma = 1 \) are the weighting parameters introduced to prioritize either CEs, WF, or OPEX in equation (4). At each timestamp in this work, 200 random sets of \( \alpha, \beta, \gamma \) were generated using Dirichlet distribution, and minimizing equation (4) with respect to
the random sets could result in a pareto set for the next-step decision making. Equation (5) is an equality constraint used to guarantee 100% execution of the computing tasks \( L_{f,T} \) within in the CTL \( k \).

Equation (6) specifies the lower and upper bound of each element in the vector \( \mathbf{w}^L \), where \( L_{DC} \) is the load capacity of the DC (ssj_ops/hour), \( \emptyset \) is a safety factor reserved for nonflexible workloads (\( L_{nf,T} \), ssj_ops/hour), and \( \sum_{p < T} L_{f,p}^T \) is the sum of the flexible workloads scheduled to time \( t \) before \( T \).

### 2.2. Case study

To demonstrate the proposed decision-making framework, we considered a large cloud DC in San Francisco, which implements the water-cooled chiller system with waterside economizer. The DC was assumed to have a maximum capacity of 100 MW [6], with the load shape follows that of a typical DC [5] (Figure 1 (a)). It was also assumed to be operated at an average server workload level of 25%, with the server mixes having a weighted average performance to power ratio as that of Figure 1 (b). In this work, we mainly focus on the case study demonstration rather than the modeling of the \( \text{PUE}_t, \text{WUE}_t, \text{t}_t^{\text{CO}_2}, \) and \( \text{t}_t^{\text{wcf}} \) (which can be predicted using either engineering or general statistical forecast models [3,7,8]), and we assumed that these parameters have been precisely predicted, as shown in Figure 1 (c), (d), (e), and (f). Additionally, the dynamic electricity rate of California Independent System Operator (CAISO) was shown in Figure 1 (g) [9], and a static California industrial water utility average (4.92 USD/kGal) [10] was used in this work.

![Figure 1](https://github.com/nuoaleon/Sustainable-data-center-load-scheduling)

Figure 1. Input parameters used for the case study demonstration: (a) DC load profile, (b) assumed weighted average performance to power ratio of the servers in the DC, (c) predicted DC PUE values, (d) predicted DC WUE values, (e) electricity carbon intensity of CAISO, (f) electricity WCF of CAISO, (g) dynamic electricity pricing of CAISO (All the input datasets and code implementations are available at https://github.com/nuoaleon/Sustainable-data-center-load-scheduling).

On the other hand, to explore the effect of the flexible workload percentages (\( p^f \)), CTL (\( k \)), and safety factor (\( \emptyset \)) on the potential benefits of intelligent load scheduling, several scenarios were designed, as summarized in Table 1. At each timestamp \( T \), the computing tasks arrive at the DC as defined in Figure 1 (a), a random fraction (\( p^f \)) of which are flexible workloads that are required to be 100% executed within \( k \) hours. In this work, we assumed that \( p^f \) and \( k \) are normally distributed with the domain restricted by \( 0 \leq p^f \leq 1 \) and \( k \geq 0 \), and large standard deviations were used to represent the high volatility of these parameters. In addition, a one-hour time step was used for the case study demonstration. It should be noted that the input parameters of this section were mainly introduced for demonstration purpose. In real applications, only DC specific inputs can be employed for reliable DC load scheduling.
Table 1. Summary of the model parameters used for scenario analysis.

| Parameters | Baseline | Low | High |
|------------|----------|-----|------|
| \( p_f \) (-) | \( \max\{N(0.4,0.2),0\}\) & \( \max\{N(0.2,0.2),0\}\) & \( \max\{N(0.6,0.2),0\}\) & \( \min\{N(0.4,0.2),1\}\) & \( \min\{N(0.2,0.2),1\}\) & \( \min\{N(0.6,0.2),1\}\) |
| \( k \) (hours) | \( \max\{N(9,3),0\}\) | \( \max\{N(6,3),0\}\) | \( \max\{N(12,3),0\}\) |
| \( \emptyset \) (-) | 0.6 | 0.5 | 0.7 |

3. Results and discussion

Figure 2 shows the corresponding DC workloads generated based on \( p_f \) of the baseline scenario, which was used as the real-time input for the framework demonstration. As discussed in Section 2, minimizing equation (4) could yield a set of potentially optimal solutions (Figure 3), depending on the weighting parameters employed to tradeoff among different objectives. The choice of the weighting parameters is typically subjective to the favor of the decision makers (e.g., an economic-oriented person is more likely to reduce the electric and water bill instead of carbon mitigation). To avoid the burden of the weighting parameters selection, we explored three extreme cases. That is, at each timestamp, the decision makers would always select either the minimal CEs, WF, or OPEX in the pareto set. And the corresponding load scheduling results were presented in Figure 4.

![Figure 2](image2.png)  
**Figure 2.** Hourly computing workloads of DC (baseline scenario).

![Figure 3](image3.png)  
**Figure 3.** Pareto set at timestamp \( T = 0 \) (baseline scenario).

![Figure 4](image4.png)  
**Figure 4.** Load scheduling results (baseline scenario): (a) based on minimal CEs in the pareto set, (b) based on minimal WF in the pareto set, (c) based on minimal OPEX in the pareto set (where part of the workloads was scheduled to the next day).

The results in Figure 4 show that the proposed decision-making framework can reasonably shift the flexible computing workloads to reduce the DC’s CEs, WF, and OPEX, which is critical for
quantitative load scheduling. The flexible workloads were mainly shifted in response to the carbon intensity, WCF, and the dynamic electricity pricing of CAISO, because the DC’s PUE and WUE values fluctuate within small ranges (Figure 1c and Figure 1d) and so does the water price. Specially, the load scheduling results are more similar for Figure 4 (a) and (b) due to characteristic of CAISO’s renewable energy production, where high penetration of renewable energy during hour 8~15 gives rise to low electricity carbon intensity and WCF. On the other hand, the load scheduling result presents multimodal in Figure 4 (c) as a result of the multimodality of the electricity rate. It should be noted that the stable PUE and WUE values for the case study demonstrated here are due to the implementation of the specific DC cooling system under the favorable weather condition of San Francisco. Thus, the load scheduling results may manifest significantly different characteristics under other situations, especially for DCs using different cooling systems under climates where large diurnal weather variation exists.

![Figure 5](image-url)  
**Figure 5.** Influence of the flexible workload percentage ($p^f$) on the reductions in the CEs, WF, and OPEX associated with flexible load shifting (optimization results were based on selecting the minimal OPEX in the pareto set at each timestamp, other optimization results are available through the link provided under Figure 1).  

![Figure 6](image-url)  
**Figure 6.** Influence of the CTL ($k$) and the safety factor ($\varnothing$) on the percent reductions in the CEs, WF, and OPEX (results were based on selecting the minimal OPEX in the pareto set at each timestamp, and $E(p^f) = 0.4$).

In addition, we explored the influence of the $p^f$, $k$, and $\varnothing$ on the intelligent load scheduling using scenarios defined in Table 1, and the results are summarized in Figure 5 and Figure 6. Two important observations arise from Figure 5. First, the savings in the CEs and the WF are nonnegligible even though the optimization results were based on selecting the minimal OPEX in the pareto set at each timestamp, which can be mainly explained by the high renewable energy production and the valley in the electricity price during hour 8~15 of CAISO. Second, the reductions in the CEs, WF, and OPEX as the result of the load shifting increase with higher flexible workload percentages. Even though it is apparent that higher levels of flexible workload could engender greater environmental and economic benefits, the CTL plays another important role in determining the potential benefits, as indicated by Figure 6, where the percent reductions in the CEs, WF, and OPEX increase with longer expected time limits. Particularly, the increases in the percent reductions in Figure 6 is nonlinear, with diminishing benefits for overlong expected time limit. Furthermore, compared with CTL ($k$), safety factor ($\varnothing$) is a
much less important factor in determining the benefits of load shifting, with the main reason being that the case study DC is under-utilized most of the time (compared to its maximum capacity). However, the influence of $\bar{\phi}$ on the load shifting potential is case-specific and could interact with $k$ or $p_f$ in determining the environmental and economic benefits. Special attention is needed when employing the proposed decision-making framework for different applications.

4. Conclusion

In this work, an integrated decision-making framework was developed for carbon-, water-, and economic-intelligent computing, which could identify environmentally and economically beneficial load shifting decisions in response to changes in the DC’s electricity and water use, and dynamics in the power grid’s carbon intensity, WCF, and electricity pricing, etc. The proposed framework could be useful for environmental policy-makings that promote the contract signings of flexible computing tasks or be applied for model predictive controls of green-conscious DCs. This work could benefit from several extended future investigations. First, the uncertainties in the predicted input parameters were not considered in this work, and future works should provide robust approaches for decision-making under such uncertainties. Second, we focused on temporal load scheduling rather than geographical load migration [11], where geographically distributed DCs could collaborate to achieve greater environmental benefits and cost reductions. However, load planning under such circumstances typically requires large-scale and centralized decisions, which poses challenges for its real-time applications. In such cases, distributed decision-making enabled by the emerging blockchain technology systems may be employed to address these challenges in future works [12].

References

[1] Masanet E, Shehabi A, Lei N, Smith S and Koomey J 2020 Recalibrating global data center energy-use estimates. Science (80-. ). 367 984–6
[2] Lei N and Masanet E 2021 Global Data Center Energy Demand and Strategies to Conserve Energy Data Center Handbook, 2nd Edition (Palo Alto, CA: WILEY)
[3] Lei N and Masanet E 2020 Statistical analysis for predicting location-specific data center PUE and its improvement potential Energy 201 117556
[4] Lei N and Masanet E 2021 Climate- and Technology-Specific PUE and WUE Predictions for U.S. Data Centers using a Physics-Based Approach
[5] JOHN J ST. 2020 Google Tests Load-Shifting at Data Centers to Capture the Grid’s Peak Clean Energy Hours | Greentech Media greentechmedia
[6] Masanet E and Lei N 2020 How Much Energy Do Data Centers Really Use? - Energy Innovation: Policy and Technology Aspen Glob. Chang. Inst. March 2020 Q. Res. Rev.
[7] Lei N 2020 A robust modeling framework for energy analysis of data centers ACM International Conference Proceeding Series (New York, NY, USA: Association for Computing Machinery) pp 177–80
[8] Liu X, Paritos P, Awalgaonkar N M, Billions I and Karava P 2018 Model predictive control under forecast uncertainty for optimal operation of buildings with integrated solar systems Sol. Energy 171 953–70
[9] Cabral L, Booth B and Peterson C 2017 California wholesale electricity prices are higher at the beginning and end of the day U.S. Energy Inf. Adm.
[10] Saralyn Bunch, Peterson J, Poehlman E, Loper S and Wilburn M 2017 Water and Wastewater Annual Price Escalation Rates for Selected Cities across the United States Dep. Energy 34
[11] Zheng J, Chien A A and Suh S 2020 Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers Joule
[12] Lei N, Masanet E and Koomey J 2021 Best practices for analyzing the direct energy use of blockchain technology systems: Review and policy recommendations Energy Policy 156 112422