Optimal Data Center Energy Management with Hybrid Quantum-Classical Multi-Cuts Benders' Decomposition Method

Zhongqi Zhao, *Student Member, IEEE,* Lei Fan, *Senior Member,* and Zhu Han, *Fellow, IEEE*

***Abstract*—The flourishing of the data era has led to a higher demand for hyper-scale data centers, resulting in a great energy gap. It is necessary to comprehensively analyze the energy management in data centers to minimize the operation cost. In this paper, we first propose a new optimal energy management model for data centers by considering energy consumers such as heating, ventilation, air conditioning (HVAC), server, solar, and battery. Then, inspired by the great computation power of quantum computing techniques, we propose a new hybrid quantum-classical multi-cuts Benders' decomposition algorithm, which utilizes quantum advantages in parallel computing for generating multi-cuts in a single iteration. Finally, experiments are conducted to verify the effectiveness and efficiency of the novel model and algorithms.**

***Index Terms*—Data Center, Optimal Energy Management, Quantum Computing, Benders' Decomposition**

NOMENCLATURE

Abbreviation

AC	Air conditioning
CT	Condense tower
DC	Data center
HVAC	Heating, ventilation, and air conditioning
LB	Lower-bound
Max	Maximum
Min	Mimimum
TL	Thermal load
UB	Upper-bound

Indices

i	Index for zones
j	Index for chillers or cooling towers
t	Index for time periods

Sets

$\mathcal{I}$	Set of zones
$\mathcal{J}^{\text{chiller}}$	Set of chillers
$\mathcal{J}^{\text{tower}}$	Set of cooling towers
$\mathcal{N}(i)$	Set of neighbor nodes of zone i
$\mathcal{T}$	Set of time periods

Parameters

β_m^{chiller}	Coefficient of mth item in chiller power
$\beta_t^{\text{e,g}}$	Electricity price of the local grid at time t
β_m^{pump}	Coefficient of mth item in pump power
β_m^{tower}	Coefficient of mth item in cooling tower power
β_m^{vent}	Coefficient of mth item in ventilation power
χ_i	Weight factor for zone i in the data center
$\dot{m}_i^{\text{Zone}}$	Air mass flow into the zone i
η^{char}	Battery charging efficiency
η^{dis}	Battery discharging efficiency
$\overline{\xi^{\text{B}}}$	Max energy requirement of battery
$\overline{p_t^{\text{chr}}}$	Max charging power to battery at time t
p_t^{dis}	Max discharging power from battery at time t
ρ^{air}	Density of air
$\theta_{i,t}$	Heat dissipation to zone i at time t
$\underline{\xi^{\text{B}}}$	Minimum energy reserve of battery
$\underline{v_t^{\text{vent}}}$	Minimum ventilation airflow speed at time t
$c^{\text{a,s}}$	Air specific heat capacity
$c^{\text{w,s}}$	Water specific heat capacity
C_i^{Zone}	Air heat capacity of Zone i
$E_{\text{inital}}^{\text{B,state}}$	Energy stored in Battery at time 0
E_t^{DC}	Energy demands by servers at time t
E_t^{misc}	Energy demands by miscellaneous at time t
E_t^{S}	Energy generated by solar system at time t
h_i	Height of data center
m_j^{chiller}	jth chiller water flow rate
m_j^{tower}	jth condense tower water flow rate
$R_{i,i'}^{\text{Zone}}$	Total resistance between adjacent zones (i, i')
S_i^{Zone}	Area of zone i
$T_{i,t}^{\text{AC},+}$	UB of zone i's AC temperature at time t
$T_{i,t}^{\text{AC},-}$	LB of zone i's AC temperature at time t
T_t^{chwr}	Return chilled water temperature t
T_t^{chws}	Supply chilled water temperature t
T_t^{conwr}	Return condense water temperature at time t
T_t^{conws}	Supply condense water temperature at time t
T_t^{out}	Outside ambient temperature at time t
$T_{i,t}^{\text{Zone,+}}$	Upper-bound of zone i's temperature at time t
$T_{i,t}^{\text{Zone,-}}$	Lower-bound of zone i's temperature at time t
$T_{i,\text{inital}}^{\text{Zone}}$	Zone temperature of zone i at time 0
v_t^{AC}	Supply airflow rate at time t
v_t^{out}	Outside ambient airflow rate at time t

Binary Decision Variables

u_t^{char}	Charging mode of battery at time t
u_t^{dis}	Discharging mode of battery at time t

This work was supported by National Science Foundation (NSF) under Grants CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, ECCS-2045978. (Corresponding author: Lei Fan.)

Zhongqi Zhao is with the Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77204 USA (e-mail: zzhao33@cougarnet.uh.edu)

Lei Fan is with the Department of Engineering Technology, University of Houston, Houston, TX 77204 USA (e-mail: lfan8@central.uh.edu).

Zhu Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701 (e-mail: zhan2@uh.edu).

$x_{j,t}^{\text{chiller}}$ Status of chiller j at time t
$x_{j,t}^{\text{tower}}$ Status of cooling tower j at time t

Integer Decision Variables

e_t^{chiller} Energy demands by all chillers at time t
e_t^{tower} Energy demands by all CTs at time t

Continuous Decision Variables

ΔE_t^{B} Battery level changes at time t
$E_t^{\text{B,state}}$ Energy stored in battery at time t
E_t^{G} Energy demands of data-center at time t
E_t^{HVAC} Energy demands by HVAC at time t
e_t^{pump} Energy demands by pump at time t
e_t^{vent} Energy demands by ventilation at time t
L_t^{heat} Total TL that needs to be removed from DC
p_t^{char} Charging power to battery at time t
p_t^{dis} Discharging power from battery at time t
$T_{i,t}^{\text{AC}}$ AC temperature of zone i at time t
$T_{i,t}^{\text{Zone}}$ Ambient air temperature of zone i at time t
v_t^{return} Returning airflow rate at time t
v_t^{vent} Ventilation airflow rate at time t

I. Introduction

With the proliferation of technologies such as cloud computing, the Internet of Things (IoT), 5G, augmented reality/virtual reality (AR/VR), and others, massive data centers have been built to meet the tremendous demand for computing resources. In data center operations, energy cost is a critical concern for the sustainable development of data centers. According to a recent study, many individual-operated data centers worldwide, such as Google in Council Bluffs, Iowa, and China Mobile in Hohhot, Inner Mongolia, consume more than 1 megawatt. As data centers are energy-intensive businesses that are anticipated to use 1% or more of the world's electricity, these trends have a significant impact on the world's energy demand and need to be investigated carefully and thoroughly [1], [2].

Concerns about running expenses at data centers have been exacerbated by recent changes in energy prices and carbon tax regulations. Data centers must manage their energy consumption wisely as the energy cost does take up a great portion of operation expenses. [3] has shown the importance of the HVAC units and IT units to the data center's energy gap. As it is illustrated in the paper, 75% of the total electric energy consumed by data centers was used by the rack critical loads (IT units), 11% by chillers, 9% by computer room air handler (CRAH) units, 1% by the lighting system, and around 4% by pumps. Stable operation is critical to data centers and the main energy source must be protected from interference.

Many papers have studied the energy management problem in data centers' battery systems, HAVC systems, and chillers. For the battery system, [4], [5] proposed an optimization to the battery with cutting-edge technology to ensure the steady functioning of the data center but also help data centers lower operating expenses in accordance with the energy distribution of time domains in the market's energy price optimization. For the HVAC system, [6] proposed a model for the HVAC system and utilized deep reinforcement learning to optimize the HVAC system operation. Ultimately, it realizes the capability to schedule data center workloads in mixed-use buildings jointly. For the chiller, [7] proposed a data center chiller system model. It adopted deep reinforcement learning to manage the electricity cost while prevent from overheating in the server zone. For the overall cost management of a data center, [8] proposed a model predictive control (MPC) formulation that contains some of the electrical load components in the data center and HVAC loads. Besides that, the model also considers solar energy and battery storage and it performs well to some extent. However, these works did not establish a comprehensive model for the energy system optimization problem of multiple HVAC equipment (e.g., chillers and cooling towers), batteries, renewable energies, and electrical loads. Accordingly, we proposed a detailed cost-driven data center energy management model in this paper that includes every component listed above.

In recent years, quantum computing (QC) has emerged as a powerful optimization tool, as it is a new paradigm for computing that has enormous potential, using quantum superposition and entanglement. QC has demonstrated quantum supremacy in problems like Grover's Algorithm [9] that allow for parallel computing. With advances in theory and manufacturing, many high-tech companies are competing in two quantum computing methods: the analog quantum model and the universal quantum gate model. In the universal quantum gate model, the quantum approximate optimization algorithm (QAOA) has been proposed and has shown promise for overcoming classical computers' barriers [10]. On the other track, papers [11] and [12] have shown decent results for job scheduling and classical optimization problems using quantum annealing. There is optimism about the potential applications of QC in the future, including machine learning, cloud computing, networking, communication, and more [13], [14]. D-Wave currently offers the quantum adiabatic annealer computer on the market with the most qubits of all the candidates. Inspired by the Ising model, D-Wave's quantum annealer computer is able to solve integer linear programming (ILP) problems by converting an ILP into a quadratic unconstrained binary optimization (QUBO) model, which depicts the energy state with coupling qubits interaction and externally applied fields.

As a result, some researchers are aware of QC's potential in solving large-scale complex systems and investigate the possibility of whether it could be applied in energy fields. It emphasizes the QC's potential in unit commitment (UC), resource planning, and load scheduling [15]. Then, as noted above, the QAOA becomes a novel resolution to the UC problem. Paper [16] introduced the hybrid quantum distributed algorithm, which is the distributed quantum surrogate Lagrangian relaxation (D-QSLR), and used it to solve a number of common UC problems. It demonstrates D-QSLR's computational effectiveness and demonstrates its enormous potential in UC optimization by observing that the algorithm maintains powerful convergence abilities and accurate output even as the system scale increases. Besides that, thanks to the researcher in [17], we can convert an NP-hard mixed integer linear programming (MILP) problem into a two-stage model to solve the question while the master problem has a shape close to ILP. This provides an opportunity for researchers in [18] to employ quantum computers in the ILP stage of solving MILP

problems. There are some pioneers who have applied this idea in the energy field. On the basis of Benders' decomposition and QAOA, paper [19] provides a hybrid quantum-classical optimization algorithm for the UC problem. The paper introduces a method for employing various cut selection criteria and tactics to control the size of the master problem (MAP) by using quantum computing to elicit and optimize a subset of cuts that will be introduced in each iteration of the Benders' decomposition (BD) scheme.

Because we recognize the importance of modeling for data centers and the powerful parallel computing capabilities of quantum computers, we developed a MILP model for a data center that is compatible with quantum computers and used a QC-assisted algorithm to investigate the possibilities of data center energy distribution and optimize operating costs. In the field of energy optimization, BD is widely employed for solving MILP problems in microgrids ([20], [21]), UC ([22], [23]). As a result, we aim to enhance the BD algorithm so that it can utilize the powerful computing capabilities of quantum computers and has better performance.

However, there are several obstacles in building the model, designing the algorithm, and simulation. The first difficulty is how to build a general linear model that covers as much data center equipment as possible. The second challenge is how to design a hybrid quantum-classical algorithm that takes fewer iterations to solve the corresponding problem above. In addition, the third difficulty is how to investigate how different multi-cuts strategies would affect the iteration of hybrid quantum-classical multi-cuts Benders' decomposition (HQCMBD). Research on this topic is currently lacking and has no predecessor. To overcome the above challenges, this paper presents a MILP model for a normal data center that contains every essential electric load, battery, and renewable energy. Especially in HVAC, for the universality of the problem, we consider each chiller and cooling tower as an individual to consider the final scheduling. For the second challenge, we overcome this issue by proposing a two-stage HQCMBD algorithm with the D-Wave quantum annealer computer. We employ quantum computers in solving the master problem, which can provide multiple feasible solutions. Then multiple subproblems are generated based on these feasible solutions. Each subproblem will return a cut for the master problem. Then we will have multi-cuts at each iteration which further leads to a speedup of the solving process. Finally, we set up some cases to investigate how different multi-cuts strategies would affect the iteration of HQCMBD. The contributions of this paper are summarized as follows:

- We propose a novel, detailed model for the data center in a MILP formulation. It contains every essential part of a data center, including electric loads, batteries, and renewable energy.
- We propose a hybrid quantum-classical multi-cuts Benders' decomposition algorithm to find the solution for the MILP problem of our data center model.
- In the context of hybrid quantum-classical multi-cuts classical Benders' decomposition, we look into iterations belonging to several multi-cuts strategies with various cases. Our experiments demonstrate that while increasing the number of cuts every iteration will result in improved iteration results, there is an ideal saddle point for the number of cuts per iteration with the defined iteration efficiency.

The rest of this paper is organized as follows: Section II introduces our system model for the data center. Section III illustrates our hybrid quantum-classical multi-cuts Benders' decomposition algorithm. Section IV first validates our algorithm and model by showing the corresponding simulation and then demonstrates and analysis how different multi-cuts strategies affect the final iteration outcomes. Finally, Section V concludes the whole paper and gives a brief overview of future work.

II. Optimal Energy Management Model

A. Data Center Energy Modeling Overview

The data center energy management system problem is to minimize the operation cost of a data center's energy consumption The data center energy management system problem is to minimize the operation cost of a data center's energy consumption within a time interval $t \in \mathcal{T}$ and $\mathcal{T} = \{1, 2, ..., T\}$. Δt is the length of time interval, and $T\Delta t$ is the full-time period that we are interested in. In addition, $t = 0$ is just a notation for the device's or ambient initial status, and it will not account for cost objective calculation. Assume E_t^{G} represents the electricity demand of the data center at time t, and $\beta_t^{\mathrm{e,g}}$ is the real-time electricity price on the main grid. The cost can be calculated as follows:

$$f(\cdot) = \sum_{t \in \mathcal{T}} \beta_t^{\mathrm{e,g}} \cdot E_t^{\mathrm{G}}, \tag{1a}$$

$$E_t^{\mathrm{G}} = E_t^{\mathrm{HVAC}} + E_t^{\mathrm{DC}} + \Delta E_t^{\mathrm{B}} - E_t^{\mathrm{S}} + E_t^{\mathrm{misc}} \ \forall t \in \mathcal{T}, \tag{1b}$$

$$E_t^{\mathrm{HVAC}} = \sum_{\mathrm{name} \in \mathrm{list}} e_t^{\mathrm{name}} \ \forall t \in \mathcal{T}, \tag{1c}$$

$$\mathrm{list} = \{\mathrm{AC}, \mathrm{vent}, \mathrm{pump}, \mathrm{chiller}, \mathrm{tower}\}.$$

Figure 1 visualizes an overview of our energy system modeling of a data center. Here, (1a) states that the total cost is a sum of the product of real-time electricity price $\beta_t^{\mathrm{e,g}}$ and external energy demand E_t^{G}. (1b) show what is external energy demand made of. (1c) list every energy consumer under E_t^{HVAC}. Besides the battery, these terms consume energy at every time interval.

- E^{DC}: Energy demand from data center operations, including energy consumed by servers, is one of the large energy demands in data center operations. It includes the energy consumed by servers for data processing, storage, transmission, and I/O operation. This term is either forecast or scheduled, and it is a known value in every time interval.
- E^{HVAC}: In (1c), multiple electrical loads make up the energy demand of the HVAC system. These fractions include energy consumed by some fans in air conditioners for producing a specific amount of airflow (e^{AC}), and other fans for delivering ventilation air (e^{vent}). In addition, cooling equipment such as water pumps (e^{pump}), chillers

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244

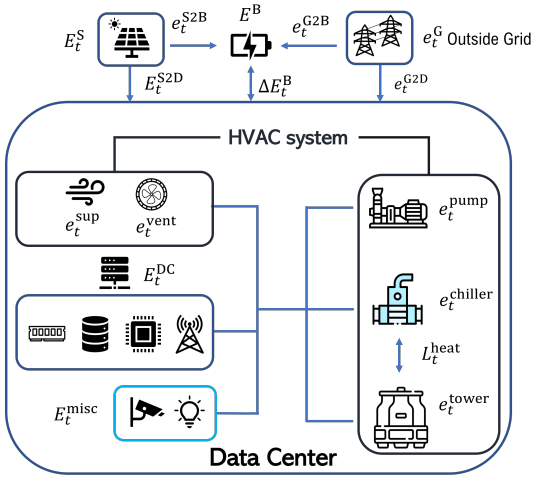


Fig. 1: Energy system model overview for a data center

(e^{chiller}) and cooling towers (e^{tower}) also contribute to the energy demand from the HVAC system. This term is varied, and it is a variable that we need to optimize within the period.

- E^{misc}: It is the energy demand from other load instances, such as lighting systems and surveillance appliances. In our model, we take it as miscellaneous demand and it is a constant number in every time interval.

Besides that, there are various electricity sources coming from both privately-owned hybrid solar systems and the main electricity grid. In addition, we use the symbol E^B to represent battery blocks connected to both the solar system and the main electricity grid.

- E^{G}: the energy provided by the main electricity grid, which belongs to the model's variables.
- E^{S}: the energy provided by the privately-owned hybrid solar system. This term is either forecast or a known value in every time interval.
- ΔE^{B}: Energy level changes of the battery energy system E^{B} owned by the data center. In this system, the data center network only purchases energy from the main grid and will not sell energy to the main grid. It belongs to the model's variables.

B. Battery Energy System

An uninterruptible power supply (UPS) is essential for data centers as it is a reliable backup or alternative energy source that can increase the system's reliability. It can fill the demand gap when local grid or solar energy input drops significantly and can also store energy when grid prices are low. The initial status of the battery energy system can be represented as

$$E_0^{\mathrm{B,state}} = E_{\mathrm{initial}}^{\mathrm{B,state}}. \tag{2}$$

1) Dynamic charging and discharging model: The charging and discharging of our battery energy system can be modeled as follows:

$$\begin{aligned} &E_t^{\mathrm{B,state}} = E_{t-1}^{\mathrm{B,state}} + \Delta E_t^{\mathrm{B}},\ \forall t \in \mathcal{T}, \\ &\text{where } \Delta E_t^{\mathrm{B}} = \left(p_t^{\mathrm{chr}} \eta^{\mathrm{chr}} - p_t^{\mathrm{dis}} \cdot (\eta^{\mathrm{dis}})^{-1}\right) \Delta t. \end{aligned} \tag{3}$$

The battery reserves will change based on the amount of charging or discharging that occurred during the previous time interval.

2) Battery reserves, discharging and charging restrictions:

$$\underline{\xi^{\mathrm{B}}} \le E_t^{\mathrm{B,state}} \le \overline{\xi^{\mathrm{B}}}, \qquad \forall t \in \mathcal{T}, \tag{4}$$

$$0 \le p_t^{\mathrm{chr}} \le \overline{p_t^{\mathrm{chr}}} \cdot u_t^{\mathrm{chr}}, \qquad \forall t \in \mathcal{T}, \tag{5}$$

$$0 \le p_t^{\mathrm{dis}} \le \overline{p_t^{\mathrm{dis}}} \cdot u_t^{\mathrm{dis}}, \qquad \forall t \in \mathcal{T}, \tag{6}$$

$$u_t^{\mathrm{dis}} + u_t^{\mathrm{chr}} \le 1, \qquad \forall t \in \mathcal{T}. \tag{7}$$

Constraint (4) represents the bounds for the energy capacity of the battery system. Constraints (5) and (6) regulate the maximum and minimum charging and discharging energy during the time interval t. Constraint (7) ensures the battery is choosing only a state from idling, charging, and discharging mode during the time interval t.

C. Space Temperature Model

The ambient temperature within the data center is important, as an abnormal temperature can damage electrical devices. Hot air can make it difficult for circuits to dissipate internal heat, and cold temperatures can cause short circuits due to frost and dew. Therefore, we need to use a dynamic temperature estimation model and impose restrictions to predict future ambient temperature. The initial temperature of every zone is represented as

$$T_{i,0}^{\mathrm{Zone}} = T_{i,\mathrm{initial}}^{\mathrm{Zone}},\ \forall i \in \mathcal{I}. \tag{8}$$

1) Dynamic Temperature Model: We use a space temperature model with multiple air conditioning units (AHUs) and calculate the energy demand of the HVAC system based on the amount of needed supply air. We adopt the idea from [24] and use an RC (resistor-capacitor) network model to estimate the future temperature of the data center zone at every time interval. We have simplified the equation from [8] and introduced a linearized dynamics model near its equilibrium operating points (temperature) as follows:

$$\begin{aligned} T_{i,t}^{\mathrm{Zone}} = T_{i,t-1}^{\mathrm{Zone}} &+ \sum_{i' \in \mathcal{N}(i)} \left(\frac{T_{i',t-1}^{\mathrm{Zone}} - T_{i,t-1}^{\mathrm{Zone}}}{C_i^{\mathrm{heat}} R_{i'i}^{\mathrm{Zone}}} \right) + \frac{\theta_{i,t}}{C_i^{\mathrm{heat}}} \\ &+ \frac{\dot{m}_{i,t}^{\mathrm{Zone}} c^{\mathrm{a,s}} \left(T_{i,t}^{\mathrm{AC}} - T_{i,t-1}^{\mathrm{Zone}}\right)}{C_i^{\mathrm{heat}}},\ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}, \\ \text{where } C_i^{\mathrm{heat}} &= c^{\mathrm{a,s}} \cdot \rho^{\mathrm{air}} \cdot S_i^{\mathrm{Zone}} \cdot h_i, \\ \dot{m}_{i,t}^{\mathrm{Zone}} &= k_i^{\mathrm{AC}} \cdot v_t^{\mathrm{AC}}. \end{aligned} \tag{9}$$

$T_{i,t}^{\mathrm{Zone}}$ is the zone's temperature at time t. Meanwhile, C_i^{heat}, and $\dot{m}_i^{\mathrm{Zone}}$ denotes the heat capacity of zone i, and the supply cooling air mass flow into zone i, respectively. k_i^{AC} is the coefficient that converts the supply cooling airflow rate to the supply cooling air mass flow into the zone and we provide a detailed function for calculating k_i^{AC} in [25]. $\theta_{i,t}$ represents the internal heat generation (e.g., heat from servers) at time t. $R_{i'i}^{\mathrm{Zone}}$ stands for the total resistance between zone i and an adjacent zone i'. C_i^{heat} is a product of the following parameters, where c_p^{air} is the specific heat capacity at room temperature;

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244

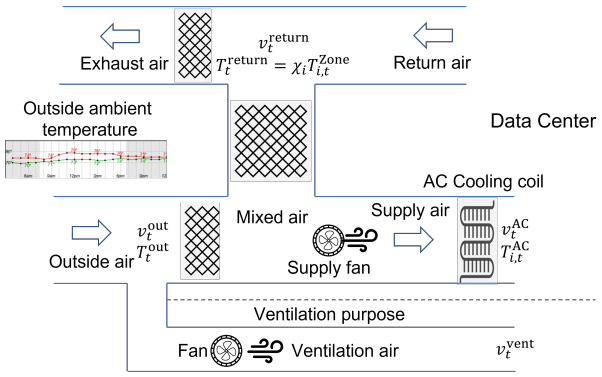


Fig. 2: Airflow demand and supply in AHUs

ρ^{air} is the density at room temperature; S_i^{Zone} is the area of each zone in the data center; h_i is the height of each zone.

2) Temperature restriction: It is important to operate a data center at a cooler ambient air temperature and to regulate the temperature of each zone within a certain range to prevent servers from overheating or reaching dew point temperature i.e.,

$$T_{i,t}^{\text{Zone},-} \leq T_{i,t}^{\text{Zone}} \leq T_{i,t}^{\text{Zone},+}, \ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}, \tag{10}$$

$$T_{i,t}^{\text{AC},-} \leq T_{i,t}^{\text{AC}} \leq T_{i,t}^{\text{AC},+}, \ \forall i \in \mathcal{I}, \forall t \in \mathcal{T}. \tag{11}$$

The previous goal can be achieved by controlling the temperature of the cooling air delivered to the corresponding zone i at time t, as shown in (11). The zone temperature range is set in (10).

D. HVAC System Model

The HVAC system is a major energy consumer in a data center. It has to meet the temperature and ventilation requirements in each data center space. In this study, we focus on HVAC cooling systems and aggregated AHUs with shared water pumps but distinct chillers and cooling towers, as opposed to the conventional approach of viewing chillers and cooling towers as a single entity [26] [27].

1) AHUs model: Figure 1 shows our airflow demand and supply model for AHUs. The AHUs use a mixture of return and outside air to cool the air before supplying it as the cooled supply air to building rooms for temperature control. In order to provide the appropriate amount of indoor airflow, the AHUs must also absorb a certain amount of outside air for ventilation purposes. We use v_t^{vent}, v_t^{out}, v_t^{return}, and v_t^{AC} to represent the ventilation airflow rate, outside (ambient) airflow rate, return airflow rate, and supply cooling airflow rate at time t. v_t^{vent} and v_t^{return} are decision variables, while the rest are parameters.

Constraints (12) and (13) enforce the requirements for ventilation and the bounds on the supply cooling airflow rate in every AHU.

$$v_t^{\text{vent}} + v_t^{\text{out}} \geq \underline{v}_t^{\text{vent}}, \ \forall t \in \mathcal{T}. \tag{12}$$

$$v_t^{\text{AC}} = v_t^{\text{out}} + v_t^{\text{return}}, \ \forall t \in \mathcal{T}. \tag{13}$$

Constraint (12) fixes the minimum ventilation airflow rate for the data center. Equation (13) states the return air and outside air are blended and cooled through AHUs before becoming the cooling air to zones. The airflow that comes out from the cooling coil is renamed as the supply cooling air v_t^{AC} to data center zones.

2) Chiller and cooling tower heating constraints: A heat exchange system between the supply air and chilled water absorbs the thermal (heat) load in the data center zone. Additionally, the cooling tower's condensed water will exhaust the thermal (heat) load. AHUs are used to process such media, and the medium circulates in a loop between the endothermic and exothermic sides. We use L_t^{heat} to represent the total thermal load that needs to be removed from the data center by our HVAC system as shown in (14). We can use constraints (15) and (16) to explain the relationship between heat exchange from three perspectives: indoor, chiller, and cooling tower.

$$\begin{aligned} L_t^{\text{heat}} = &\left(T_t^{\text{out}} - \sum_{i \in \mathcal{I}} \chi_i T_{i,t}^{\text{AC}} \right) \cdot v_t^{\text{out}} c^{\text{a,s}} \\ &+ \sum_{i \in \mathbf{I}^{\text{Zone}}} \chi_i \left(T_{i,t}^{\text{Zone}} - T_{i,t}^{\text{AC}} \right) v_t^{\text{return}} c^{\text{a,s}}, \ \forall t \in \mathcal{T}. \end{aligned} \tag{14}$$

$$\sum_{j \in \mathcal{J}^{\text{chiller}}} \alpha_{j,t}^{\text{chiller}} x_{j,t}^{\text{chiller}} \geq L_t^{\text{heat}}, \ \forall t \in \mathcal{T}, \tag{15}$$

$$\text{where } \alpha_{j,t}^{\text{chiller}} = m_{j,t}^{\text{chiller}} \left(T_t^{\text{chwr}} - T_t^{\text{chws}} \right) c^{\text{w,s}}.$$

$$\sum_{j \in \mathcal{J}^{\text{tower}}} \alpha_{j,t}^{\text{tower}} x_{j,t}^{\text{tower}} \geq L_t^{\text{heat}}, \ \forall t \in \mathcal{T}, \tag{16}$$

$$\text{where } \alpha_{j,t}^{\text{tower}} = m_{j,t}^{\text{tower}} \left(T_t^{\text{conwr}} - T_t^{\text{conws}} \right) c^{\text{w,s}}.$$

Equation (14) calculates the total thermal load L_t^{heat}. Then, constraints (15) and (16) calculate the amount of chilled water and condensed water demand for removing the internal thermal load. c_p^{air} and c_p^{water} denote the heat capacity of air and water at the operating temperature. T_t^{out} denotes the outside air temperature at time t. $T_{i,t}^{\text{Zone}}$ and $T_{i,t}^{\text{AC}}$ represent the air temperature and supply air temperature in zone i at time t, respectively. χ_i is the weight of temperature for zone i. T_t^{chws} and T_t^{chwr} are the supply and return chilled water temperatures. T_t^{conws} and T_t^{conwr} denote the supply and return condensed water temperatures. Similarly, m_t^{chiller} and m_t^{conw} represent the chilled water flow rate and the condensed water flow rate. Among all notations of temperature, $T_{i,t}^{\text{AC}}$ is the only continuous decision variable among all temperature symbols.

E. Energy Consumption

This section presents a collection of models that calculate the energy consumption of every electrical load.

1) Air conditioners' fan energy consumption: Air conditioners' fan energy consumption is a model as

$$e_t^{\text{AC}} = \beta^{\text{AC}} v_t^{\text{AC}}, \ \forall t \in \mathcal{T}. \tag{17}$$

β^{AC} is a coefficient to the air conditioner's fan airflow rate.

2) Ventilation energy consumption: Ventilation energy consumption is a model as

$$e_t^{\text{vent}} = \beta_0^{\text{vent}} + \beta_1^{\text{vent}} \left(v_t^{\text{vent}} - \underline{v}^{\text{vent}} \right), \ \forall t \in \mathcal{T}. \tag{18}$$

β_0^{vent} is a constant energy consumption for ventilation and β_1^{vent} is a coefficient to modified ventilation airflow rate.

3) Air chiller coil energy consumption: Air chiller coil energy consumption is a model as

$$e_t^{\text{chiller}} = \sum_{j \in \mathcal{J}^{\text{chiller}}} \gamma_{j,t}^{\text{chiller}} x_{j,t}^{\text{chiller}}, \ \forall t \in \mathcal{T}, \tag{19}$$
$$\text{where } \gamma_{j,t}^{\text{chiller}} = \beta_{0,j}^{\text{chiller}} + \beta_{1,j}^{\text{chiller}} m_{j,t}^{\text{chiller}}.$$

$\beta_{0,j}^{\text{chiller}}$ is a constant energy consumption for chiller j and $\beta_{1,j}^{\text{chiller}}$ is a coefficient to chiller j's coolant liquid flow rate at time t.

4) Condense tower energy consumption: Condense tower energy consumption is modeled as

$$e_t^{\text{tower}} = \sum_{j \in \mathcal{J}^{\text{tower}}} \gamma_{j,t}^{\text{tower}} x_{j,t}^{\text{tower}}, \ \forall t \in \mathcal{T}, \tag{20}$$
$$\text{where } \gamma_{j,t}^{\text{tower}} = \beta_{0,j}^{\text{tower}} + \beta_{1,j}^{\text{tower}} m_{j,t}^{\text{tower}}.$$

$\beta_{0,j}^{\text{tower}}$ is a constant energy consumption for condense tower j and $\beta_{1,j}^{\text{tower}}$ is a coefficient to tower j's coolant liquid flow rate.

5) Pump energy consumption: Pump energy consumption is modeled as

$$e_t^{\text{pump}} = \beta_0^{\text{pump}} + \beta_1^{\text{pump}} \frac{\kappa_t L_t^{\text{heat}}}{c^{\text{a,s}}}, \ \forall t \in \mathcal{T}, \tag{21}$$
$$\text{where } \kappa_t = \frac{c^{\text{a,s}}}{\left(T_t^{\text{chwr}} - T_t^{\text{chws}}\right) \cdot c^{\text{w,s}}}.$$

β_0^{pump} is a constant energy consumption for operating pumps and β_1^{pump} is a coefficient to coolant flow rate in demand.

F. Abstract Energy Management Model

Now, our problem formulation can be presented as follows:

$$\min_{\mathbf{z},\mathbf{y}} \sum_{t \in \mathcal{T}} \beta_t^{\text{e,g}} \cdot E_t^{\text{G}} \quad \text{s.t. } (1b), (1c), (2) - (21). \tag{22}$$

We rewrite the above model in an abstract energy management model as follows:

$$\min_{\mathbf{z},\mathbf{y}} \mathbf{c}^\intercal \mathbf{z} + \mathbf{d}^\intercal \mathbf{y} \tag{23a}$$
$$\text{s.t. } \mathbf{A}_1 \mathbf{z} + \mathbf{G}_1 \mathbf{y} \geq \mathbf{b}_1, \tag{23b}$$
$$\mathbf{A}_2 \mathbf{z} + \mathbf{G}_2 \mathbf{y} = \mathbf{b}_2, \tag{23c}$$
$$\mathbf{A}_3 \mathbf{z} \geq \mathbf{b}_3, \tag{23d}$$

where the binary symbol $\mathbf{z} = \left(u_t^{\text{dis}}, u_t^{\text{chr}}, x_{j,t}^{\text{chiller}}, x_{j,t}^{\text{tower}}\right)$, and the continuous symbol $\mathbf{y} = \left(p_t^{\text{dis}},\ p_t^{\text{chr}}, T_{i,t}^{\text{Zone}}, T_{i,t}^{\text{AC}}, v_t^{\text{vent}}\right)$. This problem formulation (23) is considered a MILP problem. Here, $\mathbf{z}$ is the binary variable vector, and $\mathbf{A}_n$ is its corresponding coefficient matrix in constraints. $\mathbf{y}$ is a continuous variable vector and $\mathbf{G}_n$ is a non-zero matrix and its corresponding coefficient matrix in constraints. $\mathbf{b}_n$ is the corresponding right-hand side constant vector. $\mathbf{c}^\intercal$ is the coefficient vector of variable $\mathbf{z}$ in the objective function (1a). Similarly, $\mathbf{d}^\intercal$ is the coefficient vector of variable $\mathbf{y}$ in the objective function.

For the constraint part, (23d) contains constraints where only binary variables are involved, and it is made up of (7). On the other hand, (23b) contains inequality constraints where both binary and continuous variables are involved, and it is obtained by (4)-(6), (10)-(12), and (14)-(16). Similarly, the rest of the constraints contribute to (23b) and only contain equality constraints where both binary and continuous variables are involved. The abstract energy management model is a kind of MILP problem that belongs to NP-hard problems which are difficult for classical solvers in practice when the problem goes large and complex.

III. HYBRID QUANTUM-CLASSICAL MULTI-CUTS BENDER'S DECOMPOSITION

A. Benders' Decomposition

We first introduce Benders' decomposition, a method for solving MILP problems. It involves expressing the original MILP problem, such as (23), as a master problem and a corresponding subproblem. This is done by introducing the continuous real number variable λ to represent the connection between the two problems.

(MAP) Master Problem:

$$\min_{\mathbf{z},\lambda} \quad \mathbf{c}^\intercal \mathbf{z} + \lambda \tag{24a}$$
$$\text{s.t.} \quad \mathbf{A}_3 \mathbf{z} \geq \mathbf{b}_3, \tag{24b}$$
$$\left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal u^k \geq \lambda \quad \forall\ k \in \hat{K}, \tag{24c}$$
$$\left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal r^j \geq 0 \quad \forall\ j \in \hat{J}, \tag{24d}$$
$$\mathbf{b}^* = [\mathbf{b}_1, \mathbf{b}_2]^\intercal, \mathbf{A}^* = [\mathbf{A}_1, \mathbf{A}_2]^\intercal,$$
$$\lambda \in \mathbb{R}.$$

We denote $\hat{K}$ and $\hat{J}$ as the set of extreme points $\{u^k\}$ and rays $\{r^k\}$ accumulated by the polyhedron $O = \left\{w \in \mathbb{R}_+^m \mid \mathbf{G}^* w \geq \mathbf{d}\right\}$ of the subproblem in every turn we have gone through so far. The subproblem is as follows:

(SUB) Subproblem:

$$\lambda(\mathbf{z}) = \min_{w} \ \left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal w \tag{25a}$$
$$\text{s.t.} \quad \mathbf{G}^* w \geq \mathbf{d}, \tag{25b}$$
$$\mathbf{b}^* = [\mathbf{b}_1, \mathbf{b}_2]^\intercal, \mathbf{A}^* = [\mathbf{A}_1, \mathbf{A}_2]^\intercal, \tag{25c}$$
$$\mathbf{G}^* = [\mathbf{G}_1, \mathbf{G}_2], w \in \mathbb{R}_+^m. \tag{25d}$$

In the subproblem, if the inner product between $(\mathbf{b} - \mathbf{A}\mathbf{z})$ and any dual ray $r^{j'}$ is negative, the dual problem of (25) is infeasible and $\lambda(\mathbf{z}) = -\infty$. Therefore, to determine a future direction, we generate a new feasibility cut for the master problem (24), i.e.,

$$\left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal r^{j'} \geq 0, \quad \hat{J}^* = \hat{J} \cup \{j'\}. \tag{26}$$

If x satisfies (26), then we obtain an extreme point $u^{k'}$ and the value of $\lambda(x)$ is given by

$$\begin{aligned} \lambda(x) &= \min_{k' \in K} \ \left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal u^{k'} \\ \Rightarrow &\left(\mathbf{b}^* - \mathbf{A}^* \mathbf{z}\right)^\intercal u^{k'} \geq \lambda, \quad \hat{K}^* = \hat{K} \cup \{k'\}. \end{aligned} \tag{27}$$

Equation (27) generates a new optimality cut in the master problem. In addition to the extreme points and extreme rays, we use $\hat{K}$ and $\hat{J}$ to denote the current known extreme points and extreme rays of O, respectively. For the next iteration, we update the set of extreme points u^k and rays r^k by setting $\hat{K} = \hat{K}^*$ and $\hat{J} = \hat{J}^*$.

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244

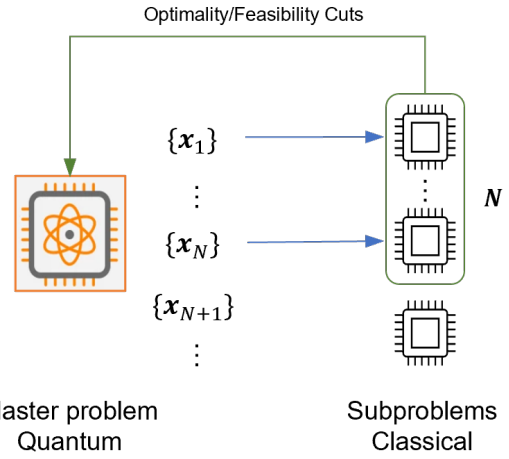


Fig. 3: An overview of multi-cuts strategy assisted by the quantum computer

B. Hybrid Quantum-Classical Benders' Decomposition

Hybrid quantum-classical Benders'r decomposition has been shown to be a powerful tool for solving MILP problems [13], [18]. To solve the master problem in (24), we use a quantum annealing computer, as it has a special formulation that can be converted into an ILP problem. The subproblem, on the other hand, is solved using a classical computer, as quantum annealing computers have difficulty solving a complete continuous linear programming model, whereas classical computers can handle it efficiently.

C. Constraints to QUBO Equivalent Penalty Pairs

The focus of this paper is a MILP problem that has constraints. While the master problem (24) has a similar structure to an ILP problem, it is not in a formulation that can be processed by a quantum annealer computer, known as a QUBO formulation. One way to transform a constrained ILP into the corresponding unconstrained QUBO form is to use penalties. We then find the optimal solution by determining the best penalty coefficients for the constraints.

D. QUBO Formulation

Quantum annealer computers are able to solve unconstrained optimization problems in a QUBO formulation. To take advantage of state-of-the-art quantum annealers offered by D-Wave, the ILP problem must be converted to the corresponding QUBO formulation. Following the steps in [18], the quantum formulation of our problem formulation is as follows:

$$f_{\text{QUBO}}(\mathbf{x}) = \mathbf{x}^\intercal \mathbf{Q}_{\text{obj}} \mathbf{x} + \mathbf{x}^\intercal \mathbf{Q}_{\text{cons}} \mathbf{x}. \tag{28}$$

The first item in (28) is the QUBO formulation of the objective function in (24a) and the second item is the QUBO formulation of constraints (24b), (24c), and (24d).

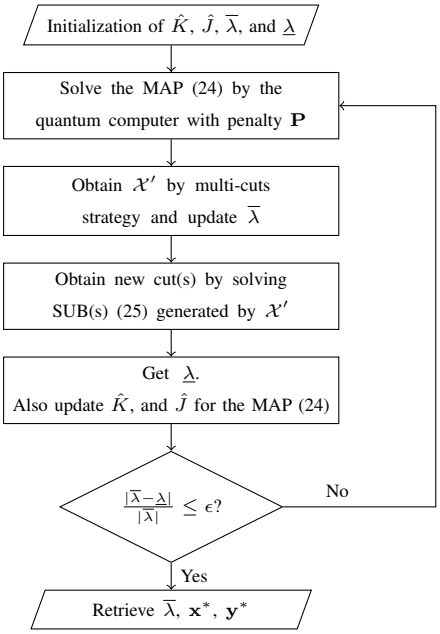


Fig. 4: Flowchart of the HQCMBD algorithm

Then, the objective function of the master problem (MAP) (24a) can be translated to a QUBO model as

$$\mathbf{x}^\intercal \mathbf{Q}_{\text{obj}} \mathbf{x} = \mathbf{z}^\intercal \text{diag}\,(\mathbf{c})\,\mathbf{z} + \lambda(\mathbf{w}) \tag{29a}$$

$$\text{where } \lambda(\mathbf{w}) = \sum_{i=0}^{m_1} w_i 2^{i-\underline{m}} w_i - \sum_{j=m_2}^{M} w_j 2^{j-m_2} w_j, \tag{29b}$$

$$m_1 = \underline{m} + \overline{m}_+, \tag{29c}$$

$$m_2 = 1 + \underline{m} + \overline{m}_+, \tag{29d}$$

$$M = 1 + \underline{m} + \overline{m}_+ + \overline{m}_-. \tag{29e}$$

(29) specifies how we build $\mathbf{x}^\intercal \mathbf{Q}_{\text{obj}} \mathbf{x}$. We introduce a binary vector $\mathbf{w}$ with a length of M bits to replace and recover the continuous variable λ in (24). Among all notations, $\overline{m}_+ + 1$ is the number of bits that are used to represent the positive integer part of the variable. $\underline{m}$ is the number of bits that are used to represent the positive decimal part of the variable, and $\overline{m}_- + 1$ is the number of bits that are used to represent the negative integer part of the variable.

$$\begin{aligned} &\mathbf{x}^\intercal \mathbf{Q}_{\text{cons}} \mathbf{x} \\ =&f^{(24b)}_{\text{QUBO}}(\mathbf{z}, \mathbf{s}) + f^{(24c)}_{\text{QUBO}}(\mathbf{z}, \mathbf{w}, \mathbf{s}) + f^{(24d)}_{\text{QUBO}}(\mathbf{z}, \mathbf{s}). \end{aligned} \tag{30}$$

(30) specifies how we build $\mathbf{x}^\intercal \mathbf{Q}_{\text{cons}} \mathbf{x}$. $f^{(i)}_{\text{QUBO}}(\mathbf{x})$ is the corresponding constraint penalty, which is made of the decision variable vector $\mathbf{x}$. In addition, we adopt a slack variable vector $\mathbf{s}$ to solve (30). [18] gives a detailed approach to turning constraints into corresponding QUBO formulations.

In short, f_{QUBO}:$\{0,1\}^n \rightarrow \mathbb{R}$ is a quadratic polynomial over binary variable vector $\mathbf{x}$. In (24), $\mathbf{x}$ contains variables of $\mathbf{z}$, $\mathbf{w}$, and $\mathbf{s}$. For the quantum annealer computers, the QUBO solver tends to find a binary vector $\mathbf{x}^*$ that minimizes f_{QUBO} among all other binary vectors. For the matrix, $\mathbf{Q}$ can be either an upper-triangular matrix or a symmetric matrix, and the cost coefficient for $x_i x_j$ is represented by q_{ij}.

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244

Algorithm 1 Hybrid Quantum-classical Multi-cuts Benders' Decomposition Algorithm

Require: Initial (Empty) sets of extreme points $\hat{K}$ & rays $\hat{J}$

$\overline{\lambda} \leftarrow +\infty$, $\underline{\lambda} \leftarrow -\infty$
while $\frac{|\overline{\lambda}-\underline{\lambda}|}{|\overline{\lambda}|} \geq \epsilon$ **do**
  $\mathbf{P} \leftarrow$ Appropriate penalties numbers or arrays
  $\mathbf{Q} \leftarrow$ Reformulate both objective and constraints in (24) and construct the QUBO formulation by using corresponding rules
  $\mathcal{X}' = \{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_N\} \leftarrow$ Solve the (24) by the quantum computer. We pick at most N feasible solutions with the lowest energies from all readings
  $\overline{\lambda} \leftarrow$ Extract $\mathbf{w}$ and replace the $\overline{\lambda}$ with $\overline{\lambda}(\mathbf{w})$
  for $\mathbf{x} \in \mathcal{X}'$ **do**
    $\lambda(\mathbf{x}) \leftarrow$ Solve problem (25) by Gurobi with classical computers
    $\underline{\lambda} \leftarrow z_{LP}(\mathbf{x})$
    if $z_{LP}(\mathbf{x}) = -\infty$ **then**
      An extreme ray j of O has been found
      $\hat{J} = \hat{J} \cup \{j\}$
    else if $z_{LP}(\mathbf{x}) < \overline{\lambda}$ **and** $\overline{\lambda} \neq +\infty$ **then**
      An extreme point k of O has been found
      $\hat{K} = \hat{K} \cup \{k\}$
      break
return $\overline{\lambda}$, $\mathbf{x}^*$, $\mathbf{y}^*$

E. Multi-cuts Strategy

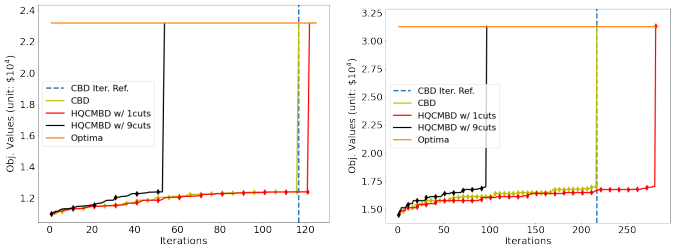
Figure 3 is an overview of our multi-cuts strategy. One advantage of quantum computing is that it can generate multiple feasible solutions from solving the master problem (MAP) (24) in each iteration. This feature allows the hybrid quantum process to potentially speed up the convergence of the hybrid quantum-classical Benders' decomposition compared to the classical computer. Our algorithm selects the top N feasible solutions with the lowest energies from the quantum reading results in each iteration. We use them as seeds to generate feasibility and optimal cuts for the next iteration on the classical computer.

F. Algorithm Framework

After we have introduced all the important components in Section III, then, the entire workflow could be described as a framework in Figure 4, and we summarize our HQCMBD method in detail in Algorithm 1. In general, the HQCMBD algorithm consists of 4 parts:

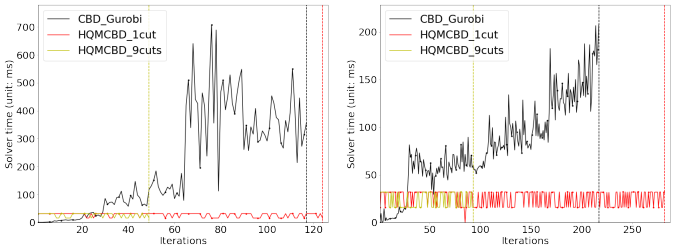
1) Initialize all necessary parameters and sets;
2) Solve the master problem (MAP) and obtain at most N solutions with lowest energies on quantum computers and update $\overline{\lambda}$;
3) Solve the subproblem (SUB) on classical computers and obtain the corresponding feasibility and optimality cuts for the master problem (MAP) in the next round and update $\underline{\lambda}$;
4) If the threshold $\frac{|\overline{\lambda}-\underline{\lambda}|}{|\overline{\lambda}|} > \epsilon$, then go to Step 2. Otherwise, terminate and return the optimal solution of the master problem (MAP) and subproblem (SUB). This step means the iterative process will get to convergence and it will be reached until the λ's gap reaches the tolerance.

This algorithm employs both quantum and classical computers to accelerate the BD method and the multi-cuts strategy is easy to modify to deal with other types of MILP problems.



(a) Case 1 (b) Case 2

Fig. 5: Objective function value of each iteration for different HQCMBD multi-cuts strategies in different case setups



(a) Case 1 (b) Case 2

Fig. 6: Solver access time of each iteration of the HQCMBD and the CBD approach in different case setups

IV. Experiment

A. Implementation Details

Both the HQCMBD algorithm and classical BD (CBD) were implemented in Python 3.8. Every iteration of the classical MILP problems was solved by the Gurobi 9.1.2 solver on a workstation with AMD 3900X processors (12 cores, 4 GHz) and 32 GB RAM. The BD master problem (MAP) instances were solved by the D-Wave hybrid quantum computer, which has over 5,000 qubits and 35,000 couplers based on the Pegasus topology. However, due to the high cost of QPU utilization and time limitations for the developer, we ran test cases that could be solved in fewer than 200 iterations for the HQCMBD since increasing numbers of slack variables, which are introduced from the objective and optimality/feasibility cuts, will make it too complicated to map itself to Pegasus topology in a Dwave quantum annealer computer [28].

B. Simulation Setup

Our data setup is available on Github [25]. To make the model more realistic, we introduced randomness to the variables $T_{\text{init}}^{\text{Zone}}$, $\beta_t^{\text{e,g}}$, S_i^{Zone}, $m_{j,t}^{\text{chiller}}$, $m_{j,t}^{\text{tower}}$, and E_t^{S}. The grid price profile $\beta_t^{\text{e,g}}$ follows a normal distribution $\mathcal{N}(\mu, \sigma)$, where μ is the average electricity rate and σ is the uncertainty of the on-grid market price. The other variables above are drawn from a uniform distribution within their respective ranges.

Table I shows that we evaluate both approaches through two different cases. The elements in the set-up column are

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244

TABLE I: Iteration comparison between HQCMBD with different multi-cuts strategies

	Set-up	$\|\mathbf{x}\|$	Iter. of CBD	Aver. iter. of HQCMBD ($N=3$)	Gain	Iter. of HQCMBD ($N=6$)			Aver. iter. of HQCMBD ($N=6$)	Gain	Aver. iter. of HQCMBD ($N=9$)	Gain
Case 1	$\{3,4,5\}$	33	117	83.67	-28%	66	74	65	68.33	-42%	56	-52%
Case 2	$\{4,2,2\}$	24	217	160	-26%	120	125	127	127.33	-41%	100	-54%

TABLE II: Operating cost comparison between models

	Set-up	Opt. Cost of model [6]	Opt. Cost Our MILP model	Saving
Case 1	$\{3,4,5\}$	\$ 25,291.621	\$ 23,147.714	8.48%
Case 2	$\{4,2,2\}$	\$ 34,843.638	\$ 30,990.856	11.06%

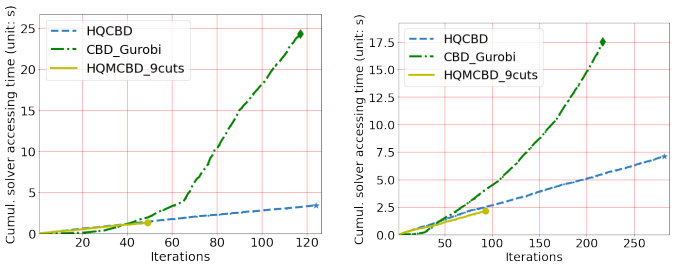
$\{|\mathcal{T}|, \mathcal{J}^{\text{chiller}}, \mathcal{J}^{\text{tower}}\}$. $|\mathcal{T}|$ represents the number of time intervals, while $\mathcal{J}^{\text{chiller}}$ and $\mathcal{J}^{\text{tower}}$ represent the number of chillers and condense towers, respectively. The first case has 3 time intervals, 4 chillers, and 5 condense towers; the second has 4 time intervals, 2 chillers, and 2 condense towers. Then, $|\mathbf{x}|$ represents the cardinality of every case's binary decision variables set. The fourth column states how many iterations CBD needs to reach the optimal solution. We will use this column as our reference to evaluate our proposed algorithm. After that, columns 5, 10, and 12 provide the average number of iterations for HQCMBD to solve different problems under different multi-cut strategies, while columns 6, 11, and 13 show the progress of the average number of iterations of HQCMBD compared to CBD. To reduce randomness and prevent bias, we ran each algorithm three times. Columns 7 to 9 in Table I show the number of iterations of the three-time test of the HQCMD 6-cut algorithm. The time interval was set to $\Delta t = 15$ minutes, a common time interval in energy management. In all experiments, the algorithm terminates itself and returns the result if $\left(\overline{\lambda} - \underline{\lambda}\right) / \overline{\lambda} < 0.1\%$.

C. Model Comparison

Compared to the model proposed in [6], our model (23) takes chillers and condense towers as individual devices rather than recognizing them as one device. The constraints of (19), (20), (15), and (16) provide detailed modeling for the cooling system, which is closer to a realistic data center. As shown in Table II, we calculate the optimal operating cost through two different models. Our MILP model's operating cost is lower than its counterpart in both cases, with improvements of 8.48% and 11.06%, respectively. It demonstrates that our detailed model not only depicts a realistic data center system that contains every essential part of a data center but also reduces the operating cost significantly compared to the model in [6].

D. Simulation Result

In this subsection, we present the results of our numerical experiments. Since both the CBD approach and the HQCMBD approach had a 100% success rate in finding the global optima in all test cases, we focus on the average number of iterations, robustness, time consumption, and multi-cuts strategy of the HQCMBD approach.



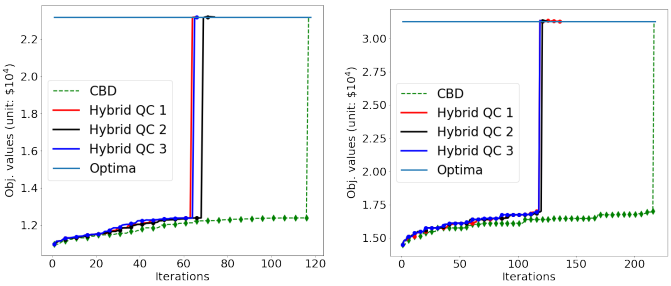
(a) Case 1 (b) Case 2

Fig. 7: Cumulative solver access time of the HQCMBD compared to the CBD approach in different case setups

1) Average Iteration Rounds: We used the data from the 9-cuts HQCMBD strategy as a comparison to CBD. Figure 5 shows 2 graphs, each representing a case in our test bench. The x-ticks are the iteration numbers, while the y-axis is the corresponding objective function value with a unit of dollars. The curves show how the objective value of the master problem changes during the iterations of both the CBD and HQCMCBD approaches. We used four different colors to label the different algorithms. Among all the poly-lines, there is one for the objective function of CBD, one for Hybrid quantum-classical (1-cut) Benders' Decomposition (HQCBD), and a line belongs to the HQCMBD algorithm. In addition, we added the optima line to indicate the position of the problem's optimal. It is clear that there are differences in the performance of the three approaches. Although the HQCBD does not beat the CBD in terms of iteration number due to the quantum randomness in generating optimal solutions from the master problem (MAP), the HQCMBD does a great job that is beyond the CBD. In each set-up and test shot, the HQCMBD outperforms the CBD with fewer iterations to reach the optima. Table I provides a detailed comparison of the number of iterations required by the classical BD and HQCMBD algorithms to solve different cases. As the problem size increases, both CBD and HQCMBD take more iterations to reach the optima. Although the number of iterations for HQCMBD varies during different tests, the average number of iterations is about 40% less than that of CBD. Therefore, thanks to its multi-cuts strategy, HQCMBD outperforms CBD in terms of iteration.

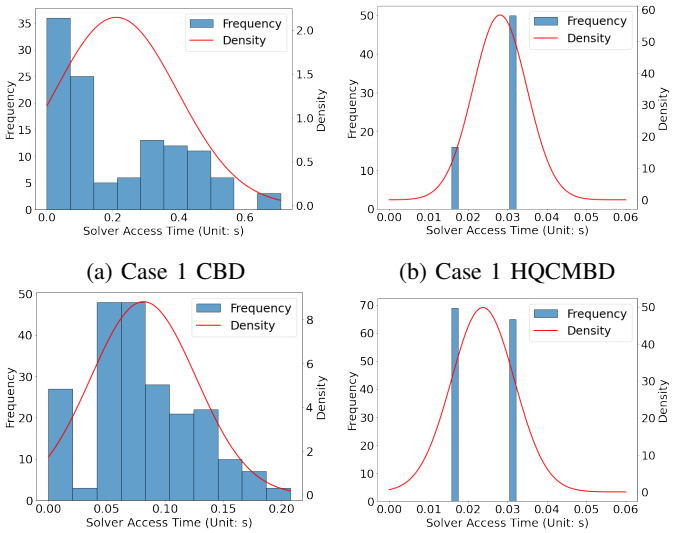
2) Time consumption: In Figure 6, we pick 3 typical methods, which are CBD, HQCBD, and 9-cuts HQCMBD, to assist our time consumption analysis. The x-ticks are the iteration numbers, while the y-axis is each iteration's corresponding solver access time with a unit of seconds. We use black, red, and yellow to represent CBD, HQCBD, and HQCMBD. For the HQCMBD, apparently, as long as the master goes complicated enough, the HQCMBD will beat the CBD in

This article has been accepted for publication in IEEE Transactions on Sustainable Energy. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2023.3309244



(a) Case 1 (b) Case 2

Fig. 8: The objective value comparison between the 6-cuts HQCMBD and the CBD approach in different case setups



(a) Case 1 CBD (b) Case 1 HQCMBD

(c) Case 2 CBD (d) Case 2 HQCMBD

Fig. 9: Solver access time histogram of the 6-cuts HQCMBD and CBD approach in different case setups

processor access time at the mid and late stages. Then, Figure 7 provides each approach's total solver access time. Although Figure 6 and 7 share the same x-ticks property, the y-axis of 7 is the cumulative access time of the corresponding solver over the current iteration round. Thanks to the less iteration and less processor access time in the mid and late stages, the HQCMBD approach shows a huge advantage over the CBD approach. Moreover, the HQCBD also benefits from the feature of less access time. Although it takes more rounds to reach the optima, it consumes less processor access time. Therefore, we can conclude that HQCMBD outperforms CBD regarding time consumption when encountering complex problems.

3) Robustness: Figure 8 shows 2 graphs, each representing a case in our test bench. The x-ticks are the iteration numbers, while the y-axis is the corresponding objective function value with a unit of dollars. Among all the poly-lines, there is one for the objective function of the CBD, and the rest belong to the three attempts of the HQCMBD algorithm. The average of three attempts' iteration with the same preset will mitigate the randomness inside the HQCMBD's master problem (MAP). Therefore, we denote those attempts as QC1, QC2, and QC3. In addition, we added the optima line to indicate the position of the problem's optimal. As shown in Table I and Figure 8, we used the data from the 6-cuts HQCMBD strategy compared to CBD. Moreover, the number of iterations required by the HQCMBD approach does not vary too much between tests. The deviation is less than 10%, which is relatively small. Regarding the multi-cuts strategy, the comparison of columns in Table I shows that the gain from the corresponding strategy does not vary much in different cases.

Besides the number of iterations, we also investigate the robustness of solver access time by using CBD and 6-cuts HQCMBD strategy. Figure 9 shows two pairs of graphs, and each graph represents the solver access time histogram outcome of a specific algorithm under a case. In the histogram figures, both sub-graphs share an identical axis layout. The x-axis has 10 bins with the automatic time interval. The left y-axis is the frequency of the solver access time within the corresponding time interval. The right y-axis is the density probability after fitting the result to bins of certain time intervals. The curve is the density curve of the solver access time's frequency. The solver access time of the CBD has a distribution that scatters wider than the HQCMBD approach.

TABLE III: Standard Deviation Comparison

Detail / Model	Standard Deviation Unit: 10^{-3}	Gain
Case1 CBD	186.0	96.33%
Case1 HQCMBD	6.8	
Case2 CBD	45.2	82.31%
Case2 HQCMBD	8.0	

Table III states the HQCMBD's standard deviation has a gain of at least 82.31% (5.6 times more robust) over the CBD approach, which implies that the former method is more robust in terms of the solver access time. What's more, it means the HQCMBD's computation performance is not sensitive to the changes in problem settings. Therefore, we can claim the robustness of the HQCMBD approach.

4) Multi-cuts strategy: We compare the performance of single-cut and multi-cuts strategies: 1 and 9-maximum-cuts per iteration. As shown in Figure 5, the x-ticks are the iteration numbers, while the y-axis is the corresponding objective function value with a unit of dollars. The number of iterations increases as the problem size increases. But the HQCMBD approach performs better with more maximum cuts.

Figure 10 shows the gain progress of each strategy. The slope of the gain progress becomes less steep as the number of maximum cuts increases. It suggests a saddle point in the multi-cuts strategy, beyond which the benefits of adding more cuts diminish. However, no explicit formula connects the saddle point to the problem size. The best strategy is problem dependent and it may change as the problem size increases. The HQCMBD generally outperforms the CBD approach, and a good multi-cuts strategy can provide significant benefits.

5) Analysis: The HQCMBD approach performs better than the CBD approach in data center energy management for several reasons. First, the HQCMBD approach has a 100% success rate, ensuring good performance. Second, the average

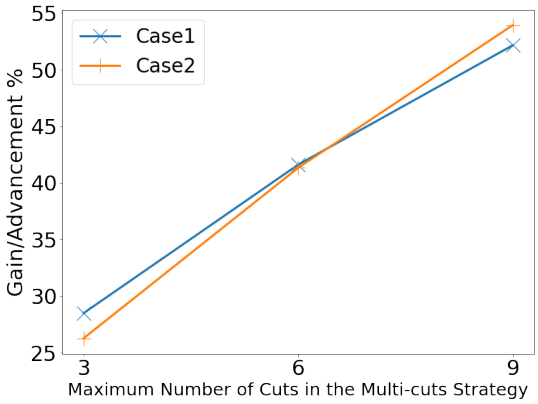


Fig. 10: Gain comparison between M-cuts HQCMBD

number of iterations for the HQCMBD approach is much less than that for the CBD approach, resulting in faster convergence to the optimal solution. Third, the HQCMBD approach is highly robust in handling different scenarios.

V. Conclusion

In this paper, we propose a MILP model for data center energy management. We then introduce the HQCMBD approach, which combines quantum and classical computers to solve the MILP model for data center energy management. Our simulation study shows that the HQCMBD approach outperforms the CBD approach regarding success rate, average iteration rounds, robustness, and multi-cuts strategy. From a solution quality perspective, the HQCMBD approach is able to converge and return correct optimal results similar to the classical algorithm. Furthermore, our approach demonstrates robustness with fewer iterations. Overall, we conclude that the proposed HQCMBD approach is a promising tool for solving MILP problems in data center energy management. What's more, it is a great demonstration of QC's computational potential and will propel us to investigate QC's application in other energy-related fields.

References

[1] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in *Proceedings of the ACM SIGCOMM conference on Data communication*, vol. 39, no. 4, New York, NY, USA, Aug. 2009, p. 123–134.

[2] A. Shehabi, S. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, and W. Lintner, "United states data center energy usage report," *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775 Page*, vol. 4, Jun. 2016.

[3] T. Xu and S. Greenberg, "Data center energy benchmarking: Part 4 - case study on a computer-testing center (no. 21)," Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Tech. Rep., Aug. 2007. [Online]. Available: https://www.osti.gov/biblio/926298

[4] C. C. Thompson, P. K. Oikonomou, A. H. Etemadi, and V. J. Sorger, "Optimization of data center battery storage investments for microgrid cost savings, emissions reduction, and reliability enhancement," *IEEE Transactions on Industry Applications*, vol. 52, no. 3, pp. 2053–2060, Jan. 2016.

[5] B. Zhang and E. N. Senior, "Distributed redundant integration of data center battery storage with the grid for regulation services," in *IEEE Power & Energy Society General Meeting (PESGM)*, Virtual, Jul. 2021.

[6] T. Wei, S. Ren, and Q. Zhu, "Deep reinforcement learning for joint datacenter and HVAC load control in distributed mixed-use buildings," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 3, pp. 370–384, Apr. 2019.

[7] Y. Li, Y. Wen, D. Tao, and K. Guan, "Transforming cooling optimization for green data center via deep reinforcement learning," *IEEE transactions on cybernetics*, vol. 50, no. 5, pp. 2002–2013, Jul. 2019.

[8] T. Wei, M. A. Islam, S. Ren, and Q. Zhu, "Co-scheduling of datacenter and HVAC loads in mixed-use buildings," in *Seventh International Green and Sustainable Computing Conference (IGSC)*, Hangzhou, China, Apr. 2016.

[9] L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, ser. STOC, New York, NY, Jul. 1996, p. 212–219.

[10] G. Guerreschi and A. Matsuura, "QAOA for max-cut requires hundreds of qubits for quantum speed-up," *Scientific reports*, vol. 9, no. 1, pp. 6903–6903, May 2019.

[11] W. Xuan, Z. Zhao, L. Fan, and Z. Han, "Minimizing delay in network function visualization with quantum computing," in *IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, Los Alamitos, CA, Oct. 2021, pp. 108–116.

[12] I. Dunning, S. Gupta, and J. Silberholz, "What works best when? a systematic evaluation of heuristics for max-cut and qubo," *INFORMS Journal on Computing*, vol. 30, no. 3, pp. 608–624, Oct. 2018.

[13] L. Fan and Z. Han, "Hybrid quantum-classical computing for future network optimization," *IEEE Network*, vol. 36, no. 5, pp. 72–76, Nov. 2022.

[14] W. O'Quinn and S. Mao, "Quantum machine learning: Recent advances and outlook," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 126–131, Apr. 2020.

[15] A. Giani and Z. Eldredge, "Quantum computing opportunities in renewable energy," *SN Computer Science*, vol. 2, no. 5, p. 393, Jul. 2021.

[16] F. Feng, P. Zhang, M. A. Bragin, and Y. Zhou, "Novel resolution of unit commitment problems through quantum surrogate lagrangian relaxation," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2460–2471, May 2023.

[17] A. M. Geoffrion, "Generalized Benders decomposition," *Journal of Optimization Theory and Applications*, vol. 10, no. 4, pp. 237–260, Oct. 1972.

[18] Z. Zhao, L. Fan, and Z. Han, "Hybrid quantum Benders' decomposition for mixed-integer linear programming," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Austin, TX, May 2022, pp. 2536–2540.

[19] N. G. Paterakis, "Hybrid quantum-classical multi-cut Benders approach with a power system application," *Computers & Chemical Engineering*, vol. 172, p. 108161, Jan. 2023.

[20] R. Jamalzadeh and M. Hong, "Microgrid optimal power flow using the generalized Benders decomposition approach," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 4, pp. 2050–2064, Oct. 2018.

[21] Z. Li and M. Shahidehpour, "Privacy-preserving collaborative operation of networked microgrids with the local utility grid based on enhanced benders decomposition," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2638–2651, Dec. 2019.

[22] A. Nasri, S. J. Kazempour, A. J. Conejo, and M. Ghandhari, "Network-constrained AC unit commitment under uncertainty: A Benders' decomposition approach," *IEEE transactions on power systems*, vol. 31, no. 1, pp. 412–422, Mar. 2015.

[23] Y. Guo and C. Zhao, "Islanding-aware robust energy management for microgrids," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1301–1309, Jun. 2016.

[24] M. Maasoumy, A. Pinto, and A. Sangiovanni-Vincentelli, "Model-based hierarchical optimal control design for HVAC systems," in *Dynamic Systems and Control Conference*, vol. 54754, Arlington, VA, May 2012, pp. 271–278.

[25] Z. Zhao. HQCMBD. [Online]. Available: https://github.com/djzts/HQCMBD

[26] J. Lyu, S. Zhang, H. Cheng, K. Yuan, Y. Song, and S. Fang, "Optimal sizing of energy station in the multienergy system integrated with data center," *IEEE Transactions on Industry Applications*, vol. 57, no. 2, pp. 1222–1234, Jan. 2021.

[27] M. H. Beitelmal and C. D. Patel, "Model-based approach for optimizing a data center centralized cooling system," *Hewlett-Packard (HP) Lab Technical Report*, Apr. 2006.

[28] S. Zbinden, A. Bärtschi, H. Djidjev, and S. Eidenbenz, "Embedding algorithms for quantum annealers with chimera and pegasus connection topologies," in *High Performance Computing: 35th International Conference, ISC High Performance 2020, Frankfurt/Main, Germany, June 22–25, 2020, Proceedings*. Springer, Jun. 2020, pp. 187–206.

Zhongqi Zhao (S'21) received a B.S. degree in electronic engineering from Beijing Jiaotong University, in 2018, a B.S. degree in Mathematics from the University of Minnesota, Twin Cities, in 2018, and B.S. and M.S. degrees in electrical engineering from the University of Minnesota, Twin Cities, in 2018 and 2020, respectively. Zhongqi Zhao is currently a Ph.D. student at the University of Houston. His research interests include quantum computing, optimization methods, complex system operations, power system operations, and planning.

Lei Fan (M'15-SM'20) is an Assistant Professor with the Engineering Technology Department at the University of Houston. Before this position, he worked in the power industry for several years. He received the Ph.D. degree in operations research from the Industrial and System Engineering Department at the University of Florida. His research includes quantum computing, optimization methods, complex system operations, power system operations, and planning.

Zhu Han (S'01–M'04-SM'09-F'14) received a B.S. degree in electronic engineering from Tsinghua University, in 1997, and M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Han's main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019, and ACM distinguished Member since 2019. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Technical Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."