Co-Scheduling of Datacenter and HVAC Loads in Mixed-Use Buildings

Tianshu Wei, Mohammad Atiqul Islam, Shaolei Ren, Qi Zhu
University of California, Riverside
{twei002, misla006, shaolei, qizhu}@ucr.edu

***Abstract*—The majority of datacenters are within mixed-use facilities, where they often share some common infrastructures and energy supplies with other operations (e.g., non-IT offices and labs). In such mixed-use buildings, two major energy loads are datacenter IT equipment and HVAC (heating, ventilating, and air conditioning) system. The HVAC demand comes from both datacenter rooms and other non-IT rooms. To effectively lower peak demand and reduce energy cost for mixed-use buildings, it is important to leverage the scheduling flexibility from both the HVAC system and the delay-tolerant datacenter workload in a *collaborative* fashion. In this work, we model the major physical and cyber components of mixed-use buildings, and propose a model predictive control (MPC) formulation to co-schedule datacenter and HVAC loads, with consideration of solar energy and battery storage. The MPC formulation minimizes building energy cost while satisfying various requirements on room temperature, ventilation, and datacenter workload deadlines. Compared with separate scheduling strategy, our approach significantly reduces peak demand and overall energy cost, and provides better leverage of renewable energy supply. Furthermore, we demonstrate that our formulation is also effective in reducing carbon footprint, and balancing its trade-off with energy cost.**

I. Introduction

Building stock is energy-intensive and consumes 40% of the global electricity usage [20]. In particular, with the massive growth of power-hungry datacenters supporting the exploding digital economy, *mixed-use buildings* (MUBs), which include both datacenter operations and significant space for other usage (e.g., offices) [19], have emerged as one of the most significant energy consumers.

Datacenters in MUBs are highly diverse, ranging from state-of-the-art commercial datacenter (e.g., Equinix [5]) to large scientific computing clusters and to small-/medium-size server rooms. Recent studies have shown that "the majority of datacenters are physically located within mixed-use facilities" [19]. In fact, mega-scale dedicated datacenters only take up around 4% of the total datacenter energy consumption, whereas the remaining 96% goes to other types of datacenters that are mostly located in MUBs [13]. While the space for datacenter in an MUB may not be dominant, its energy demand could be large due to high power density (0.1-1kW per square foot [17]), e.g., a real-world measurement shows that datacenter load accounts for approximately 50% of an MUB's overall energy demand [1]. It is estimated that the combined energy usage by MUBs with datacenter operations are responsible for 4% or more of the worldwide electricity consumption, with a projected quick growth to over 6% by 2020 [13].

While MUBs with datacenters are prevalent and have huge power demands, optimizing their energy management has not been sufficiently addressed. Existing efforts on datacenter energy efficiency, albeit encouraging, mostly focus on dedicated datacenters (e.g., Google), where all the space and supporting infrastructure (e.g., cooling and electrical systems) are for datacenter operations [19], [3]. On the other hand, the vast literature on building energy efficiency mostly focuses on non-datacenter load management such as HVAC control [9], [14], while treating datacenters as "miscellaneous" plug-in loads and ignoring the high scheduling flexibilities of many datacenter workloads. Some recent studies began to holistically manage HVAC load with energy storage for cost savings [22], [21]. Nonetheless, these studies are not applicable to MUBs with datacenter loads, which have their own unique dynamics determined by IT workload arrivals and scheduling decisions and require a different and also much richer set of control knobs (e.g., servers turned on/off, workload deferment). Moreover, shared HVAC components (e.g., chillers) between datacenter rooms and office rooms present new challenges in modeling and co-management.

The lack of coordination between datacenter load management and non-datacenter load management (e.g., HVAC control) often results in energy inefficiency. For instance, MUBs may have unnecessarily high peak power demand, if datacenter and non-datacenter loads are not carefully managed to avoid peaking their individual power demands at the same time. Furthermore, the two types of loads typically share the limited on-site renewable energy supplies. Without coordinating their consumption, the overall energy usage may not be able to effectively follow the availability of renewables and leverage it for energy cost reduction.

In view of the critically important but little-investigated MUBs with datacenters, we extend the building energy management literature by uniquely incorporating and leveraging the large yet flexible datacenter loads that offer new cost saving opportunities. We propose a novel energy management approach that schedules the energy demands of datacenter and non-datacenter operations in a *coordinated fashion*[1]. More specifically, ***the main contributions of our work*** are:

- We model the major physical and cyber components in MUBs, as shown in Fig. 1, which include datacenter work-

[1]While there are various types of MUBs, we focus on those that are managed by a single building manager, e.g., enterprise MUBs that house office occupants and in-house private datacenters.

978-1-5090-5117-5/16/$31.00

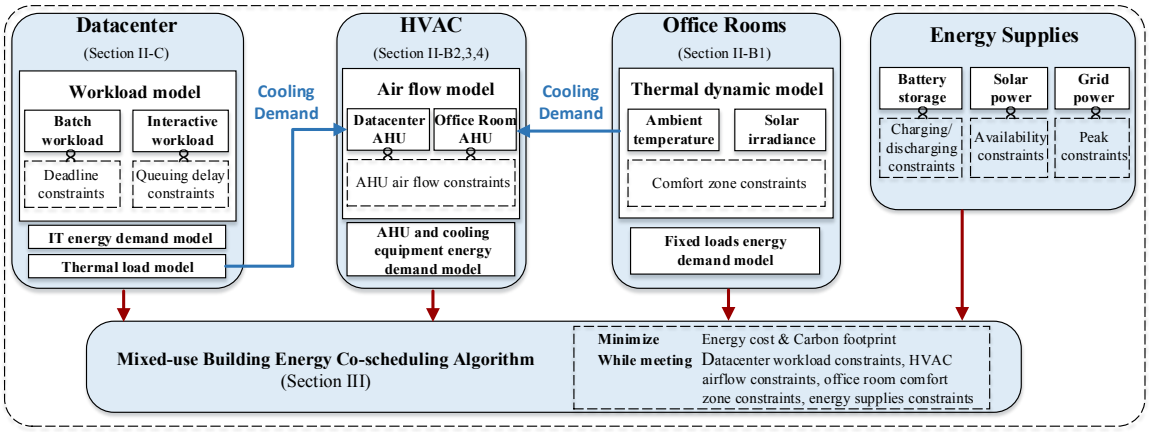


Fig. 1: Modeling of cyber and physical components and MPC-based co-scheduling algorithm for mixed-use buildings

load and energy/thermal demand of IT equipment, office room thermal dynamics and comfort zone constraints on room temperature, and HVAC operation and energy demand for both datacenter and office rooms. These models set the foundation for our co-scheduling approach.

- We develop a model predictive control (MPC) based formulation for co-scheduling datacenter and HVAC loads, with the consideration of intermittent renewables and battery storage, to minimize total energy cost while satisfying requirements on temperature, ventilation and datacenter workload deadlines. The formulation can also be extended for minimizing carbon footprint, an important environmental objective, and evaluating its trade-off with energy cost.
- We demonstrate the effectiveness of our co-scheduling approach through real-world trace-based simulations, showing that our approach may provide up to 5%-17% reduction in energy cost when compared with a separate scheduling policy and is also effective in reducing carbon footprint.

The rest of the paper is organized as follows. Section II introduces our modeling of MUBs with datacenter. Section III presents our co-scheduling formulation. Section IV shows experimental results. Section V concludes the paper.

II. System Modeling

A. MUB Energy Modeling Overview

Fig. 2 presents an overview of our modeling of the three major energy loads in an MUB with datacenter.

- e_{IT}: energy demand from datacenter operations, including energy consumed by servers for data processing (e_{server}), by uninterruptible power supplies (e_{ups}), and by power distribution units (e_{pdu});
- e_h: energy demand from HVAC system, including energy consumed by air handling units (AHUs) in office rooms (which further includes $e_{fan,o}$ for delivering supply air and $e_{vent,o}$ for ventilation), by AHUs in datacenter rooms (which includes $e_{fan,dc}$ and $e_{vent,dc}$), and by shared cooling equipment such as water pump (e_{pump}), chiller ($e_{chiller}$) and cooling tower (e_{tower});
- e_m: energy demand from other miscellaneous loads that are assumed to be fixed in our model, including lighting system, office appliances, etc.

On the energy supply side, we consider energy provided by the power grid (e_g), by renewable sources (in particular solar energy e_r), and by battery storage (e_b). The battery storage system stores energy either from power grid during off-peak hours (e_{g2b}) or from excessive renewable sources (e_{r2b}), to help shave building's peak demand.

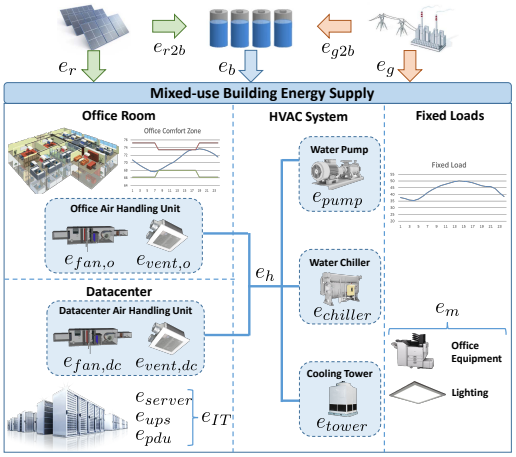


Fig. 2: Energy modeling overview for MUBs

B. Office Rooms and HVAC System Modeling

HVAC system is a major energy consumer in MUBs, and responsible for meeting the temperature and ventilation requirements in both office rooms and datacenter rooms. In this work, we consider HVAC cooling systems with separate AHUs for office and for datacenter, and with shared water pump, chiller and cooling tower[2].

[2]There are also systems with separate chillers, which can be addressed as a special case in our formulation.

Fig. 3 shows our model of air flow demand and supply in AHUs. The AHUs take a mixed of outside air and return air [6], cool it through cooling coil with chilled water, and serve it as supply air to building rooms for maintaining temperature. In addition, the AHUs need to input certain amount of outside air for ventilation purpose to ensure acceptable indoor air quality, as required by the ASHRAE standard [2].

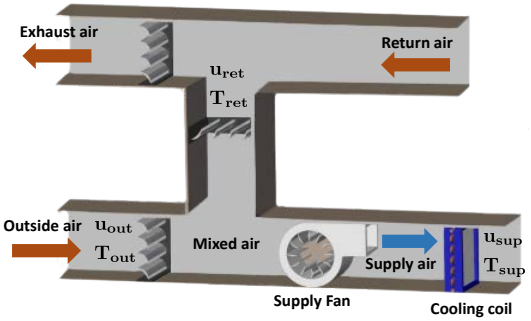


Fig. 3: Air flow demand and supply in air handling units

Next, we will first introduce our room temperature model under cooling supply air from the AHUs, and then present how we calculate the HVAC system energy demand based on the amount of needed supply air.

1) Office room temperature model: We use an RC (resistor-capacitor) network model to estimate the office room temperature changes, similarly as in literature [10]. The network consists of wall nodes and room nodes. The temperature change of a room node i is captured in Equation (1), where T_{r_i} , C_{r_i} and $\dot{m}_{r_i}$ denote the temperature, heat capacity and air mass flow into the room i, respectively. R'_{ij} stands for the total resistance between room i and adjacent node j. c_a is the specific heat capacity of air. A_{ω_i} is the total area of window on walls surrounding room i. τ_{ω_i} is the transmissivity of glass of window in room i. $\omega_i = 0$ indicates that room i does not have any window, while $\omega_i = 1$ otherwise. q''_{rad_i} is the radiative heat flux density radiated to room i, and $\dot{q}_{int_i}$ denotes the internal heat generation (e.g., heat from human occupancy). $\mathcal{N}_{r_i}$ is the set of all of the neighboring nodes to room i.

$$C_{r_i}\frac{dT_{r_i}}{dt} = \sum_{j\in\mathcal{N}_{r_i}} \frac{T_j - T_{r_i}}{R'_{ij}} + \dot{m}_{r_i}c_a(T_{s_i} - T_{r_i}) + \omega_i\tau_{\omega_i}A_{\omega_i}q''_{rad_i} + \dot{q}_{int_i} \quad (1)$$

Similarly, the temperature change of a wall node is represented by Equation (2), where T_{ω_i}, C_{ω_i}, α_i and A_i represent the temperature, heat capacity, heat absorption coefficient and area of i-th wall, respectively. Radiative heat flux density on wall i is denoted by q''_{rad_i}. $\mathcal{N}_{\omega_i}$ is the set of all of neighboring nodes to node wall i. $r_i = 0$ indicates wall i is a peripheral wall and $r_i = 0$ otherwise. More details of the model can be found in [10].

$$C_{\omega_i}\frac{dT_{\omega_i}}{dt} = \sum_{j\in\mathcal{N}_{\omega_i}} \frac{T_j - T_{\omega_i}}{R'_{ij}} + r_i\alpha_i A_i q''_{rad_i} \quad (2)$$

Differential equations of room and wall nodes can be transformed into the following state space Equation (3).

$$\dot{\boldsymbol{x}}_{\boldsymbol{t}} = f(\boldsymbol{x}_{\boldsymbol{t}}, \boldsymbol{u}_{\boldsymbol{t}}, \boldsymbol{d}_{\boldsymbol{t}}); \quad \boldsymbol{y}_{\boldsymbol{t}} = \boldsymbol{C}\boldsymbol{x}_{\boldsymbol{t}} \quad (3)$$

where $\boldsymbol{x_t}$ is the state vector representing the temperature of each node (in this work, we use bold notation to denote a vector or matrix). $\boldsymbol{u_t}$ denotes the air mass flow into each room (corresponding to $\dot{m}_{r_i}$ in Equation (1)). $\boldsymbol{d_t}$ captures the environment disturbance. Finally, $\boldsymbol{y_t}$ represents the temperature of each room node and is calculated out of system state $\boldsymbol{x_t}$.

We use the nonlinear model in Equation (3) as the plant model to estimate the actual temperature evolution in our simulation. While for efficient control in our MPC-based formulation, we use the following linear representation, derived by linearizing the nonlinear dynamics model in (3) near its equilibrium operating points:

$$\boldsymbol{x}_{\boldsymbol{t+1}} = \boldsymbol{A}\boldsymbol{x}_{\boldsymbol{t}} + \boldsymbol{B}\boldsymbol{u}_{\boldsymbol{t}} + \boldsymbol{E}\boldsymbol{d}_{\boldsymbol{t}}; \quad \boldsymbol{y}_{\boldsymbol{t}} = \boldsymbol{C}\boldsymbol{x}_{\boldsymbol{t}} \quad (4)$$

where $\boldsymbol{A}$ is the system state coefficient matrix, $\boldsymbol{B}$ and $\boldsymbol{C}$ are control and output matrices respectively, and $\boldsymbol{E}$ integrates the impacts of various environmental factors.

For office rooms, we use Equation (4) to determine the needed air mass flow $\boldsymbol{u_t}$ for maintaining room temperature. For datacenter rooms, we need a different model for calculating the air flow rate, as shown later in Section II-C.

2) Air flow constraints: For both office rooms and datacenter rooms, following constraints set the requirements on ventilation and the bounds on supply air flow rate in AHUs. $\boldsymbol{u_{vent}}(t)$, $\boldsymbol{u_{out}}(t)$, $\boldsymbol{u_{ret}}(t)$ and $\boldsymbol{u_{sup}}(t)$ represent the ventilation air flow rate, outside air flow rate, return air flow rate, and supply air flow rate at time t (more strictly speaking, during *time interval* t in our discrete control model), respectively.

$$\boldsymbol{u_{vent}}(t) + \boldsymbol{u_{out}}(t) \geq \boldsymbol{U_{vent}} \quad (5)$$

$$\boldsymbol{u_{sup}}(t) = \boldsymbol{u_{out}}(t) + \boldsymbol{u_{ret}}(t) \quad (6)$$

$$\boldsymbol{U}^- \leq \boldsymbol{u_{sup}}(t) \leq \boldsymbol{U}^+ \quad (7)$$

Constraint (5) sets the minimum ventilation requirements $\boldsymbol{U_{vent}}$ for rooms that are served by each AHU. Constraint (6) represents the fact that the outside air and the return air are mixed and cooled through AHUs to generate the supply cooling air to rooms, as shown in Fig. 3. $\boldsymbol{u_{sup}}(t)$ is determined by the total air mass flow input into rooms that are served by each AHU. Constraint (7) sets the lower and upper bounds for the supply air flow rate.

3) Chilled water and condense water constraints: The thermal (heat) load in the building is first removed via heat exchange between supply air and chilled water that flows through AHUs. Then, condense water circulates in the loop between the chiller and the cooling tower to further release buildings' thermal load to the outside environment.

Equation (8) calculates the total thermal load $L_{heat}(t)$ that needs to be removed from the building by the HVAC system (through cooling the supply air for both office rooms and datacenter rooms). c_p^{air} denotes the heat capacity of air. $\boldsymbol{T_{out}}(t)$, $\boldsymbol{T_{ret}}(t)$ and $\boldsymbol{T_{sup}}$ denote the outside air temperature, return

air temperature and supply air temperature, respectively. They are given as parameters in this work. Then, constraints (9) and (10) calculate the amount of chilled water and condense water needed for removing the building thermal load. T_{chws} and T_{chwr} denote the supply and the return chilled water temperature, respectively. T_{cws} and T_{cwr} denote the supply and the return condense water temperature, respectively. $m_{chw}(t)$ and $m_{cw}(t)$ denote the chilled water flow rate and the condense water flow rate, respectively.

$$L_{heat}(t) = (\boldsymbol{T_{out}}(t) - \boldsymbol{T_{sup}})^T \cdot \boldsymbol{u_{out}}(t) \cdot c_p^{air} + (\boldsymbol{T_{ret}}(t) - \boldsymbol{T_{sup}})^T \cdot \boldsymbol{u_{ret}}(t) \cdot c_p^{air} \quad (8)$$

$$(T_{chwr} - T_{chws}) \cdot m_{chw}(t) \cdot c_p^{water} = L_{heat}(t) \quad (9)$$

$$(T_{cwr} - T_{cws}) \cdot m_{cw}(t) \cdot c_p^{water} = L_{heat}(t) \quad (10)$$

4) HVAC system energy demand: The energy demand of HVAC system includes the energy demand from chiller, water pump, cooling tower, and fans for cooling and ventilating purpose, as modeled in below.

$$e_{sup,o}(t) = \beta_1 \sum [u_{sup,o}^{(i)}(t)]^3, \ i \in \text{office fans} \quad (11)$$

$$e_{vent,o}(t) = \beta_1 \sum [u_{vent,o}^{(i)}(t)]^3, \ i \in \text{office fans} \quad (12)$$

$$e_{sup,dc}(t) = \beta_2 \sum [u_{sup,dc}^{(i)}(t)]^3, \ i \in \text{datacenter fans} \quad (13)$$

$$e_{vent,dc}(t) = \beta_2 \sum [u_{vent,dc}^{(i)}(t)]^3, \ i \in \text{datacenter fans} \quad (14)$$

$$e_{chiller}(t) = a_0 + a_1 m_{chw}(t) + a_2 m_{chw}(t)^2 + a_3 m_{chw}(t)^3 \quad (15)$$

$$e_{pump}(t) = b_0 + b_1 m_{chw}(t) + b_2 m_{chw}(t) + b_3 m_{chw}(t) \quad (16)$$

$$e_{tower}(t) = c_3 m_{cw}^3(t) \quad (17)$$

$$e_h(t) = e_{sup,o}(t) + e_{sup,dc}(t) + e_{vent,o}(t) + e_{vent,dc}(t) + e_{chiller}(t) + e_{pump}(t) + e_{tower}(t) \quad (18)$$

Equations (11) through (14) calculate the energy demand by fans for cooling and ventilation in office rooms, and by fans for cooling and ventilation in datacenter rooms, respectively, using the model in [11]. $u_{sup,o}^{(i)}(t)$ represents the total cooling air flow rate to office rooms that are served by AHU i, and $u_{vent,o}^{(i)}(t)$ denotes the total fresh air rate that is not conditioned by the cooling coil within current time interval. $u_{sup,dc}^{(i)}(t)$ and $u_{vent,dc}^{(i)}(t)$ are similarly defined. Equations (15) through (17) calculate the energy demand by chiller, water pump and cooling tower, respectively, following the model in [15]. Note that $m_{chw}(t)$ and $m_{cw}(t)$ are calculated above in Equation (9) and (10). Finally, Equation (18) calculates the total HVAC energy demand.

C. Datacenter Modeling

We now model the datacenter IT energy demand (i.e., energy directly spent on IT equipment for computation), the delay performance constraint for its workloads, and the thermal load in datacenter for removing the generated heat. Our model is consistent with the existing literature [8], [16], [12] and captures the first-order effects of workload scheduling decisions on the datacenter energy.

$$e_s(t) = [x_a(t) + x_b(t)] \cdot e_0 \quad (19)$$

$$e_{IT}(t) = e_s(t) + \alpha \cdot e_s(t) \quad (20)$$

$$\frac{1}{\mu_a - \lambda(t)/x_a(t)} \leq D \quad (21)$$

$$x_b(t) \geq \frac{\sum_{i=1}^{j} b_i(t)}{\mu_b} \quad (22)$$

$$\sum_{t=A_i}^{A_i+N_i-1} b_i(t) = B_i \ , \ \forall i = 1...j \quad (23)$$

$$x_a(t) + x_b(t) \leq X \quad (24)$$

$$u_{dc}(t) = e_{IT}(t)/[(T_{ret,dc} - T_{sup}) \cdot c_p^{air}] \quad (25)$$

IT energy demand: The energy demand of the servers $e_s(t)$ is calculated in Equation (19) based on the average energy demand of a single server (denoted as e_0) [8] and number of active servers processing interactive and batch workloads (denoted as $x_a(t)$ and $x_b(t)$, respectively). Note that the number of servers is approximated as continuous in this work, since a datacenter often houses hundreds to thousands of servers. Equation (20) calculates the total energy demand of IT equipments. It includes server energy demand $e_s(t)$, as well as energy demand of the supporting power equipments (power distribution units and uninterrupted power supplies) which is proportional to the server energy demand $e_s(t)$ and captured using a coefficient α. $e_{IT}(t)$, however, does not include the energy demand for running the AHUs in datacenter rooms, which is addressed later in Equation (25).

Workload constraints: Datacenter processes both delay-sensitive interactive workloads and delay-tolerant batch workloads. Interactive workloads (e.g., web requests) require fast responses, whereas batch workloads (e.g., MapReduce) usually only have a deadline constraint. Constraint (21) determines the number of active servers for processing the interactive workloads based on the queueing model M/M/k to meet the performance constraint D [16]. μ_a and $\lambda(t)$ denote the maximum service rate and the arrival rate of interactive workloads, respectively. Constraint (22) determines the number of active servers for processing the batch workloads, where $b_i(t)$ denotes the amount of workload processed for the i-th batch job and μ_b denotes the maximum service rate for batch workloads. Equation (23) guarantees that each batch job B_i is finished within N_i time intervals since its arrival time A_i [8]. Constraint (24) bounds the maximum number of available servers. Aligned with existing literature [8], [16], we adopt a datacenter level control for our problem where the algorithm decides the number of servers to keep "ON" with performance constraint of (21) and (22). The server level job scheduling that captures other constraints like data locality, can be considered as secondary control that takes the number of available sever as an input.

Datacenter thermal model: We consider most of the energy consumed in server zone is converted into heat load [12], and calculate the amount of discharge air needed to remove the heat generated in the datacenter rooms in Equation (25) which is then taken into account in Equation (8) for calculating the total building thermal load.

III. Co-Scheduling Formulation

Based on the models developed in Section II, we propose an online model predictive control (MPC) formulation to co-schedule the datacenter and the HVAC energy loads, with consideration of renewables and battery storage, for minimizing energy cost and satisfying operation requirements.

The MPC-based scheduling is optimized periodically. At each time interval t, a solution of control sequence is determined by minimizing the total energy cost within the current predicting window w, while meeting the building comfort and datacenter service requirements. Then, only the first entry in the control sequence (the one corresponding to time interval k) is implemented to operate building's flexible loads, i.e., control the HVAC system and the datacenter workloads. Next, the predicting window is advance by one time interval, and the MPC-based scheduling is optimized again to determine the operation for the next time interval.

Part of the MPC-based co-scheduling formulation is as below, while the rest of the formulation include Equation (5) to (25) in Section II.

$$\mathbf{min} \sum_{t=k}^{k+w-1} [p_g(t)e_g(t) + p_b e_b(t)] + \hat{p}_g[\max_{t=k}^{k+w-1}[e_g(t)] - \hat{e}_g(t)]^+/I \quad (26)$$

$\forall t \in [k, k+w-1]$ **subject to:**

$$e_g(t) = e_{IT}(t) + e_h(t) + e_m(t) + e_{g2b}(t) - e_r(t) - e_b(t) \quad (27)$$

$$e_g(t) \geq 0 \quad (28)$$

$$\boldsymbol{T_c}(t+1) = \boldsymbol{A} \cdot \boldsymbol{T_c}(t) + \boldsymbol{B} \cdot \boldsymbol{u_{office}}(t) + \boldsymbol{E} \cdot \boldsymbol{d}(t) \quad (29)$$

$$\boldsymbol{T}^-(t+1) \leq \boldsymbol{C} \cdot \boldsymbol{T_c}(t+1) \leq \boldsymbol{T}^+(t+1) \quad (30)$$

$$e_r(t) + e_{r2b}(t) \leq E_r(t); e_{r2b}(t) \geq 0; e_r(t) \geq 0 \quad (31)$$

$$0 \leq e_b(t) \leq d_r \quad (32)$$

$$0 \leq e_{r2b}(t) + e_{g2b}(t) \leq c_r \quad (33)$$

$$S(t+1) = S(t) + \rho \cdot [e_{r2b}(t) + e_{g2b}(t)] - e_b(t) \quad (34)$$

$$E^- \leq S(t+1) \leq E^+ \quad (35)$$

Objective Function: Equation (26) defines the objective function for minimizing the total energy cost within the predicting window $(k, ..., k+w-1)$. The first term of the equation calculates the total energy consumption cost, including power grid electricity cost and battery depreciation cost. $p_g(t)$ denotes the grid electricity price at time interval t and $e_g(t)$ denotes the grid electricity consumed at t. p_b denotes the battery depreciation cost and $e_b(t)$ denotes the amount of energy discharged from battery at t.

The second term addresses the peak power demand charge. $\hat{e}_g(t)$ denotes the peak energy consumption of a time interval *before* the current interval t. If the maximum energy consumption of any time interval within current predicting window exceeds $\hat{e}_g(t)$, the amount of difference will be divided by the interval length I to get average power demand and charged with a rate $\hat{p}_g$ (we use the peak power demand charge rate from utility company as $\hat{p}_g$). Then, $\hat{e}_g(t)$ will be updated to the new peak energy consumption. Note that this second term is not the actual peak power demand charge but rather a penalty to lower the peak demand during optimization. In practice, peak power demand charge is calculated based on the highest demand within a billing cycle (often a month), while the predicting window of our MPC formulation is at the granularity of hours (set to 24 hours in experiments).

Energy Demand/Supply Constraints: Equation (27) balances the energy demand and supply. That is, the total energy demand from the datacenter energy demand $e_{IT}(t)$, the HVAC system demand $e_h(t)$, the fixed load demand $e_m(t)$ and the battery charging demand $e_{g2b}(t)$, minus the renewable energy supply (solar energy in this work) $e_r(t)$ and the battery energy supply $e_b(t)$, should be equal to the grid electricity demand $e_g(t)$. Constraint (28) requires the grid electricity consumption to be non-negative, since in our model we assume the building does not inject energy back to the grid.

Note that the datacenter energy demand $e_{IT}(t)$ and the HVAC system energy consumption $e_h(t)$ are calculated in Equations (20) and (18), respectively, as defined in Section II.

Office Room Temperature Constraints: Equation (29) shows the linearized room temperature model (a simple rewriting of Equation (4)). The room temperatures in the next time interval are estimated based on the current temperatures, air mass flow input $\boldsymbol{u_{office}}(t)$, and the environmental disturbances $\boldsymbol{d}(t)$ (e.g., sun radiation intensity, human occupancy and ambient temperature). Constraint (30) ensures that the office room temperature will not violate the comfort zone requirement.

Solar Energy Constraints: When solar energy is available during daytime, it can be applied to meet the building's energy demand. Moreover, the excessive energy may be stored in the battery storage system. Constraint (31) ensures that the solar energy usage does not exceed the available solar energy $E_r(t)$ and is non-negative.

Battery Storage Constraints: Constraint (32) sets the maximum discharging rate d_r for the battery. Constraint (33) sets the maximum charging rate c_r for the battery. As shown in Fig. 2, charging energy may come from the grid (denoted by $e_{g2b}(t)$) or from the renewables (denoted by $e_{r2b}(t)$). Equation (34) updates the state of charge of the battery, denoted by $S(t)$, where ρ is the round trip efficiency. Constraint (35) ensures that the state of charge will be within a given range (for efficient battery usage).

Carbon Footprint Optimization: In addition to energy cost, carbon footprint is another important metric as many MUBs, especially those pro-sustainability MUBs, are actively seeking green certifications. Due to the heterogeneous and time-varying composition of energy sources in producing grid power (e.g., solar, nuclear and thermal power), the carbon footprint per kilowatt (i.e., carbon efficiency) may vary significantly throughout the day [24]. More importantly, carbon efficiency differs from electricity cost efficiency (e.g., coal-produced electricity is inexpensive but very carbon-intensive), and hence we need to factor carbon footprint into our co-scheduling decisions as a new metric [24]. Towards this end,

we extend the objective function in (26) to the following Equation (36) to address both carbon footprint and total energy cost. We use a weight w_c to convert carbon footprint into an equivalent monetary value to indicate the relative importance of carbon emissions, and also to trade off between carbon footprint and energy cost.

$$\min \sum_{t=k}^{k+w-1} [p_g(t)e_g(t) + p_b e_b(t) + w_c \cdot p_c(t)e_g(t)] + \hat{p}_g[\max_{t=k}^{k+w-1}[e_g(t)] - \hat{e}_g(t)]^+/I, \quad (36)$$

where $p_c(t)$ is the average carbon efficiency calculated based on the energy fuel mix at time t [24].

IV. Experimental Results

Simulation Setup: Our grid electricity price profile is a practical time-of-use tariff with three tiers of price [4] to bill business customers whose power demand is between $200KW$ and $500KW$, a typical range for small to medium MUBs. Moreover, customers' peak power demand within a month is charged at a rate of $\$16.37/KW$ (this rate is used as $\hat{p}_g$ in the objective function). In the MPC-based formulation, the time interval length I is set to one hour and the predicting window spans 24 time intervals (i.e., $w = 24$).

For datacenter, we use I/O traces from 6 RAID volumes in Microsoft Research (MSR) at Cambridge as the batch workload [26]. For interactive workloads we use traces from server usage log of Florida International University (a large public university in the U.S.) [3]. In practice, the interactive workload may be estimated based on historical workload trace. In our experiment, we vary the ratio between interactive and batch workloads, the percentage of batch workload takes up from 20% to 80%. The delay tolerance D of interactive workload is set to $50ms$. For batch workload, each job is required to be finished before its deadline (set to 24 hours in our experiments). The maximum service rate of each server is set to 100 requests per second, and the maximum power demand of each server is $0.4KW$. The office comfort temperature range is set to $20°C - 23°C$ during day time, and relaxed to $19°C - 24°C$ at night due to low occupancy activities [7].

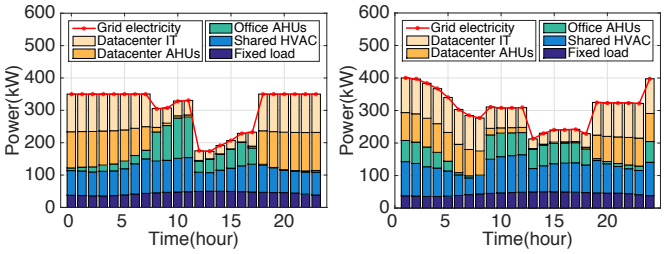
Based on Tesla's PowerWall battery storage system [18], the battery depreciation cost p_b is set to $0.09\$/KW$, and its round-trip efficiency is set to 92%. The battery capacity is set to $300KWh$, and its state of charge thresholds are 20% and 80% of its capacity, respectively. The maximum amount of charging/discharging energy in one hour is set to 25% of its capacity. The peak solar power supply during the day time is set to $150KW$, which is around 50% of the building's average power demand. The solar power is proportional to the solar radiation. We take the solar radiation data from [23] for June, 2010 (the latest available data year).

[3]We also tried another workload trace from Google's publicly available real-time traffic data [25] and the results demonstrate similar trends. We only report the results from the university traces here due to space limitation (and also because they are better representatives of datacenter workloads in MUBs).

In the following, we compare our co-scheduling approach with a baseline approach where datacenter loads and office room HVAC are scheduled separately using MPC to reduce energy cost. More specifically, for the separate scheduling approach, we divide the co-scheduling formulation introduced in Section III into two MPC-based formulations, one for scheduling datacenter operations (and the corresponding HVAC activities) and the other for scheduling office room HVAC. All experiments are simulated for one month to take into account of the monthly peak demand charge. We first study the case where 80% of requests in datacenter are batch workload, and then in section IV-D we explore other scenarios with higher interactive workload percentage.

A. Effectiveness of Co-Scheduling without Renewables and Battery

First, we conduct experiments to evaluate the effectiveness of our co-scheduling formulation, without considering renewables and battery storage. The initial estimated value of the peak energy consumption within a time interval[4], i.e., $\hat{e}_g$ in the objective function (26), is set to $350KWh$ based on the analysis of simulation data (in practice, it could be based on historical data). In the separate scheduling approach, this value is proportionally reduced based on the demands of datacenter operation and of office room HVAC control.



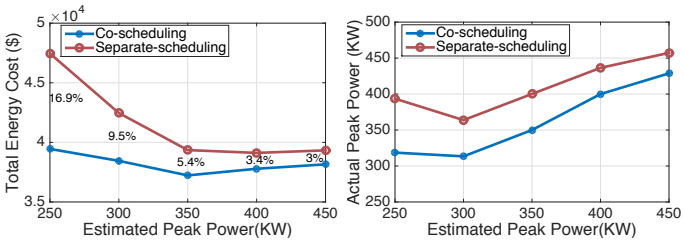
(a) Co-scheduling (b) Separate scheduling

Fig. 4: Power consumption of co-scheduling and separate scheduling without renewables and battery

Fig. 4 shows the power consumption comparison between co-scheduling and separate scheduling approaches in a weekday. Various energy demand types are represented with different colors, including datacenter IT operations, datacenter AHUs, office room AHUs, shared HAVC (chiller, water pump, cooling tower), and fixed load. The red curve shows the total grid electricity usage. From the figure, we can see that *our co-scheduling approach is more effective in reducing the energy consumption during peak hours from 12:00 to 17:00 and in reducing peak demand*. In contrast, the baseline separate scheduling approach has a higher energy consumption during peak hours and al so higher peak demand, due to the lack of coordination between datacenter load scheduling and office room HVAC control.

[4]Because the length of time interval I is set to one hour in our experiment, we will use *initial estimated peak power* interchangeably in the following sections.

Next, we vary the initial estimated peak power from $250KW$ to $450KW$ and evaluate the performance of our co-scheduling approach in different cases.



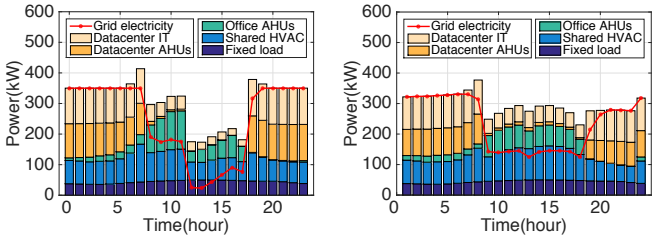
(a) Total monthly cost (b) Actual peak power demand

Fig. 5: Comparison between co-scheduling and separate scheduling in energy cost and peak power demand under various initial estimated peak power (energy cost reduction percentage is shown in the figure)

Fig. 5 shows the comparison of total monthly energy cost (including energy consumption cost and peak demand charge) and actual peak power demand between co-scheduling and separate scheduling approaches. We can see that the initial estimated peak power has significant impact on the total energy cost (in particular for separate scheduling) and on the eventual peak demand. Furthermore, in all cases, *the co-scheduling approach can significantly reduce the total energy cost and peak power demand, compared with separate scheduling.*

B. Effectiveness of Co-Scheduling in Leveraging Renewables

Next, we conduct experiments to evaluate the effectiveness of our co-scheduling approach in leveraging renewables (solar energy in this case). As stated before, we assume a solar energy profile with $150KW$ peak supply. For the separate scheduling approach, we assume the solar energy is proportionally allocated to datacenter and office rooms, based on the estimation of their demands.



(a) Co-scheduling (b) Separate scheduling

Fig. 6: Power consumption of co-scheduling and separate scheduling with solar energy supply

Fig. 6 shows the power consumption comparison between co-scheduling and separate scheduling approaches in a weekday with solar energy supply, with initial estimated peak power demand set to $350KW$. We can see that our co-scheduling approach is much more effective in leveraging the solar energy for reducing energy demand to the grid during peak hours.

Fig. 7 shows the comparison under various initial estimated peak power demand. As shown in the figure, *our co-scheduling approach can achieve a* 14.2% *cost reduction at the lowest total energy cost point, compared with the separate scheduling approach.* In addition, compared with the co-scheduling case without renewables (subsection IV-A), a 31.5% cost reduction is achieved in average.

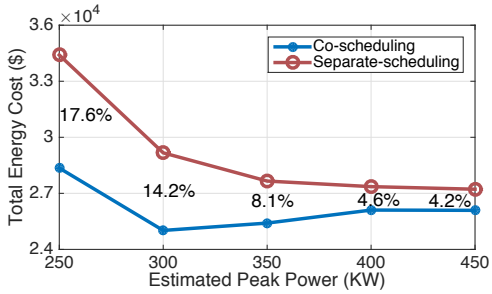
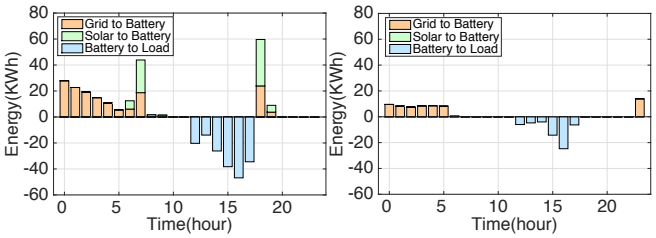


Fig. 7: Comparison between co-scheduling and separate scheduling (with solar energy supply) in energy cost (energy cost reduction percentage is shown in the figure)

C. Joint Consideration of Renewable Energy Supply and Battery Storage System

We also conduct experiments to compare co-scheduling and separate scheduling approaches with solar power (peak supply at $150KW$) and battery storage system ($300KWh$ capacity), and with initial peak power demand set to $350KW$. For separate scheduling, the battery capacity is proportionally allocated to datacenter and office rooms based on their demands.

Experimental results show that our co-scheduling approach again is more effective than the separate scheduling approach, with a 11.8% *monthly energy cost reduction.* Compared with the co-scheduling case with solar but without battery (subsection IV-B), an additional 4.1% cost reduction can be achieved.



(a) Co-scheduling (b) Separate scheduling

Fig. 8: Comparison of battery charging/discharging between co-scheduling and separate scheduling approaches

Fig. 8 shows the charging/discharging energy of battery in each time interval in both co-scheduling and separate scheduling approaches. We can see that our co-scheduling algorithm can better leverage the battery by charging the battery from grid and solar power during off-peak hours, and discharging the battery during peak hours.

D. Effectiveness of Co-Scheduling with Different Percentage of Batch Workload

We also evaluate the effectiveness of our co-scheduling approach with different ratios between interactive and batch

workload. We reduce the total amount of batch workload from 80% to 20% among all workload, and evaluate the cost reduction co-scheduling can achieve (with respect to separate scheduling) under various initial estimated peak power.

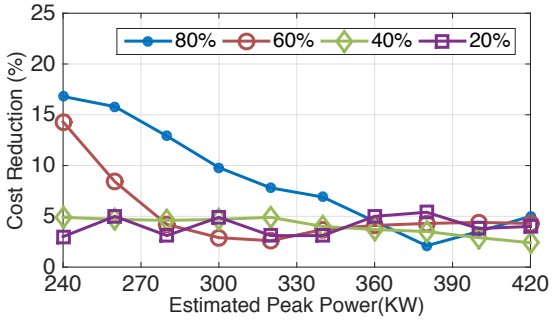


Fig. 9: Cost reduction under different batch workload ratios

As shown in Fig. 9, our co-scheduling approach can achieve more cost reduction compared to separate scheduling when there is higher level of batch workload. The performance of co-scheduling becomes insensitive to the initial estimated peak power when the percentage of batch workload decreases. That is because less scheduling flexibility can be provided by datacenter with a small fraction of batch workload.

E. Consideration of Carbon Footprint

As introduced in Section III, our co-scheduling formulation may also be used for reducing carbon footprint, using an extended objective function as shown in Equation (36). Fig. 10 shows the carbon footprint and energy cost of co-scheduling and separate scheduling approaches under different values of weight w_c. Initial estimated peak power is set to $350KW$ to study the trade-off between carbon footprint and energy cost without renewables or battery storage. From Fig. 10, we can see that the co-scheduling approach is more effective in reducing carbon footprint, at the expense of higher energy cost (due to the difference between carbon efficiency and electricity cost efficiency over time).

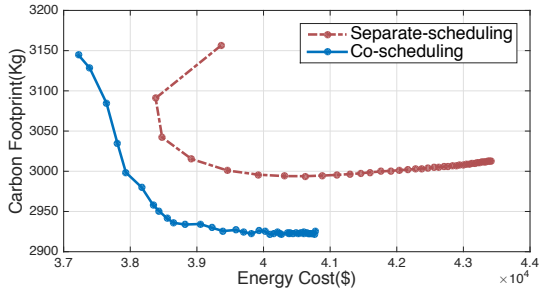


Fig. 10: Carbon footprint and energy cost of co-scheduling and separate scheduling approaches

V. Conclusions

In this work, we address the energy management of mixed-use buildings with datacenter. We propose models for major cyber and physical components in MUBs with datacenter, and propose a co-scheduling formulation to collaboratively schedule the energy demand from datacenter operations and HVAC control. Our experimental results demonstrate that our co-scheduling approach can significantly reduce energy cost and carbon footprint, when compared with a baseline approach with separate scheduling.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-1551661, CNS-1565474, ECCS-1610471, and CCF-1553757, and by the Riverside Public Utilities (Energy Innovation Grant).

References

[1] S. K. Aggarwal, L. M. Saini, and A. Kumar. Electricity price forecasting in deregulated markets: A review and evaluation. *Journal of Electrical Power & Energy Systems*, 2009.
[2] ASHRAE. Ventilation for acceptable indoor air quality. https://www.ashrae.org.
[3] L. A. Barroso, J. Clidaras, and U. Hoelzle. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool, 2013.
[4] SCE. Schedule tou-gs-3: Time-of-use, general service-demand metered. https://www.sce.com.
[5] Equinix. www.equinix.com.
[6] S. Goyal and P. Barooah. Energy-efficient control of an air handling unit for a single-zone VAV system. *CDC*, 2013.
[7] W. Liping, P. Mathew, and X. Pang. Uncertainties in energy consumption introduced by building operations and weather for a medium-size office building. *Energy and Buildings*, 2012.
[8] Z. Liu, Y. Chen, et al. Renewable and cooling aware workload management for sustainable data centers. *SIGMETRICS*, 2012.
[9] Y. Ma, F. Borrelli, et al. Model predictive control for the operation of building cooling systems. *IEEE Transactions on Control Systems Technology*, 2012.
[10] M. Maasoumy, A. Pinto and Alberto A. Sangiovanni-Vincentelli. Model-based hierarchical optimal control design for HVAC systems. *DSCC*, 2011.
[11] M. Mehdi and A. Sangiovanni-Vincentelli. Total and peak energy consumption minimization of building HVAC systems using model predictive control. *Design & Test*, 2012.
[12] R. Neil. Calculating total cooling requirements for data centers.
[13] NRDC. Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers. 2014.
[14] F. Oldewurtel, A. Parisio, et al. Energy efficient building climate control using stochastic model predictive control and weather predictions. *ACC*, 2010.
[15] H. Phillip. Model predictive control of HVAC systems: Implementation and testing at the University of California, Merced. *Lawrence Berkeley National Laboratory*, 2010.
[16] L. Rao, X. Liu, et al. Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment. *INFOCOM*, 2010.
[17] N. Rasmussen. Calculating space and power density requirements for data centers. *APC White Paper*.
[18] Tesla. http://www.teslamotors.com/powerwall.
[19] The Green Grid. Pue: A comprehensive examination of the metric, 2012.
[20] U.S. DoE. Buildings energy data book.
[21] T. Wei, T. Kim, et al. Battery management and application for energy-efficient buildings. *DAC*, 2014.
[22] T. Wei, Q. Zhu, and M. Maasoumy. Co-scheduling of HVAC control, EV charging and battery usage for building energy efficiency. *ICCAD*, 2014.
[23] NSRDB. http://rredc.nrel.gov.
[24] P.X. Gao, et al. It's not easy being green. *SIGCOMM*, 2012.
[25] Google transparency report, http://www.google.com/transparencyreport/traffic/explorer.
[26] E. Thereska, A. Donnelly, and D. Narayanan, "Sierra: a power-proportional, distributed storage system," *Tech. Rep. MSR-TR-2009-153*, 2009.