An Edge Based Data-Driven Chiller Sequencing Framework for HVAC Electricity Consumption Reduction in Commercial Buildings

Zimu Zheng* ††, Qiong Chen†, Cheng Fan‡, Nan Guan*, Arun Vishwanath§, Dan Wang*, Fangming Liu†
*The Hong Kong Polytechnic University, †Huazhong University of Science and Technology,
‡Shenzhen University, §IBM Research Australia, ††Technical Innovation Department,
Cloud BU, Huawei Technologies Co.Ltd

Abstract—It is well-known that the HVAC (heating, ventilation and air conditioning) dominates electricity consumption in commercial buildings. In this paper, we focus on one of the core problems in building operation, namely *chiller sequencing* to reduce HVAC electricity consumption. Our contributions are threefold. First, we make a case for why it is important to quantify the performance profile of a chiller, namely coefficient of performance (COP), at *run-time*, by developing a data-driven COP estimation methodology. Second, we show that predicting COP accurately is a non-trivial problem, requiring considerable computation time. To overcome this barrier, we develop a data-driven COP prediction model and an edge-based chiller sequencing framework integrating the COP predictions, and show that they strike a good balance between electricity saving and ease of use for real-world deployment. Finally, we evaluate the performance of our scheme by applying it to real-world data, spanning 4 years, obtained from multiple chillers across 3 large commercial buildings in Hong Kong. The results show an electricity saving of over 30% compared to baselines. We offer our edge based data-driven chiller sequencing framework as a cost-effective and practical mechanism to reduce electricity consumption associated with HVAC operation in commercial buildings.

1 Introduction

Centralized chilled water based HVAC plants are commonly used for cooling in large commercial buildings. These HVAC plants consume anywhere between 40% and 70% of a building's total electricity consumption [1], [2], a vast majority of which can be attributed to the chillers in the HVAC. In several nations around the world, the electricity bill paid by commercial buildings, which is dominated by the energy consumption of the HVAC, is often in the top-three list of an organization's operating expenses [3]. This trend is putting upward pressure on facility managers to improve the energy efficiency of their buildings by means of reducing the electricity consumption associated with HVAC operation.

Various techniques have been proposed in the literature for mitigating the energy impact of building HVAC. These include controlling the HVAC based on the spatio-temporal profile of occupancy inside a building [4], pre-cooling a building in advance of expected increase in occupancy [5], and incorporating renewables such as solar panels and battery storage into the energy mix [6].

While all of these approaches have merit, in this paper, we focus on the problem of chiller sequencing for reducing HVAC electricity consumption in commercial buildings. Chiller sequencing refers to operating the most efficient combination of chillers in a building at real-time to meet the time-varying cooling demand. For example, sequencing a building with two chillers [0.5, 0.7] implies that chiller 1 and chiller 2 are operating at 50% and 70% of their maximum rated capacity, respectively. Thus, the sequencing problem is to allocate the cooling load at any given time to the chillers in the most energy efficient manner so that the overall cooling demand of the building is satisfied while at the same time the electricity consumed by the chillers is kept at a minimum [7].

Most prior work on chiller sequencing has focused on developing techniques for predicting the cooling demand accurately. However, the efficacy of chiller sequencing control also relies heavily on the *run-time* performance profile of the chillers, namely the COP under different cooling load regimes. COP is a measure of the energy-efficiency of a chiller and captures the cooling power that it can output for a certain input power consumption. COP is typically greater than 1; larger values implying better efficiency [8].

Despite advances in chiller performance profiling, they have relied on developing fixed form thermodynamic models for obtaining the COP given cooling load [9]. These models have limited value in practice because COP is highly dependent on a variety of factors such as the operating conditions, configuration dynamics, varying cooling demands, degradation over time, weather and so on, making it extremely difficult to capture the impact of these parameters accurately in an analytical model. For e.g., it was recently shown that over 12 months, there was a 20% reduction in the chilled water flow rate, caused by excessive fouling that blocked the tubes in the chiller condenser [10]. Using the COP from these fixed form models therefore can introduce large errors, rendering them impractical for use in the real-world. In practice, facility managers often perform chiller sequencing using COP profiles obtained when the chillers are first tested and commissioned during installation in a building, called *initial profiles*. The initial profile considers cooling load as the sole parameter. For reasons mentioned

above, and detailed in the rest of the paper, these initial profiles fail to capture the impact of other real-world parameters, and thus are not accurate. It is evident that robust estimation of the run-time COP of chillers is critical for the success of any chiller sequencing technique.

Inspired by the advent of IoT deployment in buildings, and the availability of IoT sensor data logged by modern building management systems (BMS), in this paper, we advocate a data-driven COP profiling approach to facilitate chiller sequencing. Our COP estimation relies on data collected by BMS. Specifically, our COP profiling techniques are underpinned by BMS data obtained at 30 minute intervals from 17 chillers, over four years, across three high-rise office buildings located in Hong Kong. We make three important observations. First, existing thermodynamic models for COP estimation can be inaccurate. [For example, the run-time COP of water-cooled chillers with constant-speed primary pumps (like the ones considered in this paper) does not increase monotonically with the cooling load, as is typically assumed in practice and found in the initial profiles [11].] Second, there is a significant difference between the COP obtained from the data-driven approach and initial profiles for different cooling loads. Third, and most importantly, data-driven profiling increases the accuracy of chiller COP estimation, paving the way for energy-efficient chiller sequencing in practice. In this context, the contributions of this paper are:

- We demonstrate that there is a need for individualized COP chiller performance profiling at run-time, which when done effectively can be instrumental in reducing HVAC electricity consumption. As discussed above, the resulting COP values can vary substantially from that obtained via initial profiling. The latter is often used for sequencing chillers in practice today, undermining their energy efficiency considerably.
- We show that COP performance profiling using data-driven techniques is a challenging problem, in terms of computation time. And so we propose a data-driven COP prediction model along with an edge-based chiller sequencing control framework under time and budget constraints. We highlight that it strikes a good balance between reducing electricity consumption for chiller operation and ease of use for real-world deployment.
- We comprehensively evaluate the efficacy of our approach by applying the solution on BMS data, spanning 4 years (2012-2015), obtained from multiple chillers across 3 high-rise office buildings in Hong Kong. The results show that our chiller sequencing approach is able to save on average 21 MWh of electricity consumption in each of the 3 buildings, which is an improvement of around 30% compared to the current mode of operation of the chillers in the buildings.

Our proposed data-driven COP estimation technique and edge based chiller sequencing solution does not require any major capital expense and uses data readily available from any modern BMS. The solution recommends a chiller sequencing strategy that not only satisfies a building's cooling demand but also keeps the electricity consumption to a minimum. We offer our approach as an attractive mechanism for building facility managers to use who are on the look out for simple and low-cost means for reducing the energy and cost footprints of their buildings.

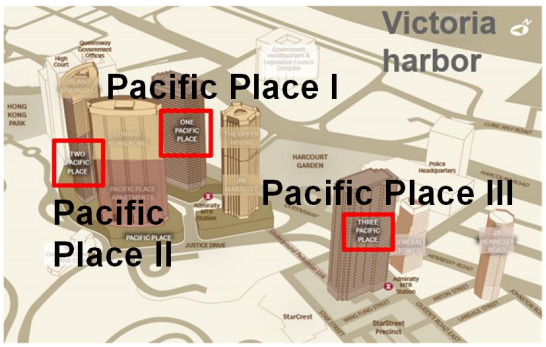


Fig. 1. Towers of Pacific Place I, II and III in Hong Kong.

2 Need for Data-driven COP Prediction for Chiller Sequencing in HVAC

2.1 Background of Chiller Plants and COP

Chiller plants are frequently used to generate cooling power for office buildings. For instance, in three office towers located in Hong Kong (Fig. 1), three chiller plants containing a total of 17 chillers serve more than ten thousand people. Four year data, spanning 2012 through 2015, at 30 minute intervals, for these different chiller models from Trane, shown in Table 1, was collected from the BMS.

TABLE 1
Chiller information in each building.

Building Name	Regular Chiller	Backup Chiller	Vendor
Pacific Place I	4 × CVHG1100	2 × CVHE370	6 × Trane
Pacific Place II	3 × CDHG2250	2 × CVHG780	5 × Trane
Pacific Place III	4 × CVGF500	2 × CVGF500	6 × Trane
Total Number	11	6	17

In commercial buildings, sequencing of chillers is performed to keep the electricity consumed for meeting a certain cooling demand to a minimum. It follows two steps, i.e. Sequence Determination and Feedback Control, and they work as follows. When a cooling demand $\mathcal{D}$ arrives, the HVAC plant needs to determine the set of chillers that need to be active and the total cooling load $Q > \mathcal{D}$ to support the demand (Sequence Determination). The HVAC plant then needs to adjust the cooling load of each (active) chiller until the cumulative load of Q is attained (Feedback Control).

COP Computation. Chiller sequencing relies heavily on the energy-efficiency of the chillers. Clearly, electricity consumption increases as a function of the cooling load. Note that the amount of electricity consumed by a chiller is not only determined by Q but also by its energy-efficiency. Intuitively, if this efficiency is low (e.g. due to poor maintenance), then more electricity will be consumed to support a required cooling demand. It is therefore of paramount importance to accurately quantify the energy-efficiency of a chiller, which is measured by its COP, determined as follows.

The cooling load of a HVAC plant at a given time is the sum of the cooling load Q_i over all chillers i, i.e., $Q = \sum_i Q_i$, where $Q_i = c_i \times m_i \times \Delta_{Ti}$. Here, c_i is the

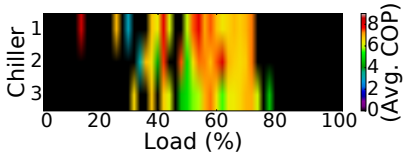


Fig. 2. Average COP for different chillers and loads.

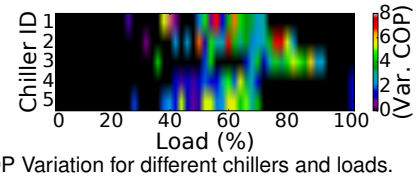


Fig. 3. COP Variation for different chillers and loads.

thermal capacity of water (kJ/kg°C), m_i is the chilled water mass flow rate (kg/s) and Δ_{Ti} is the temperature difference between the returned and supplied chilled water (°C) [12]. All these quantities are logged by our BMS.

The COP of chiller i to support cooling load Q_i is given by Q_i/E_i, where E_i is the electrical power consumed by chiller i to deliver the required amount of cooling. In practice, after a HVAC plant is installed in a building, there is a commissioning phase wherein a set of cooling loads is tested to ascertain the performance of the chillers. Following each test, values of Q and E are recorded which enable the facility manager to determine the corresponding COP profiles for the chillers. Once in production, certain statistical averaging techniques are used in the ensuing short period of operation to update the COP under different cooling loads [13], [14].

2.2 Observation of Significant COP Variation

[Reliable chiller sequencing depends on the COP across all the loading conditions for chiller i. However, when we communicated with facility managers and applied the above computation on the historical data retrieved from the BMS, we learned and confirmed that not only does the COP degrade over time and raise after maintenance, which is well-known, but the COP fluctuates markedly over different cooling loads and environmental conditions.]

To be specific, we first plot the average COPs as a function of the cooling loads in Fig. 2. We picked the first regular chiller from each building. It can be seen that these chillers often operate between 40% and 80% load, and their COP fluctuates somewhat randomly between zero and eight. The COP values for other loads are missing. For the same chillers, we plot in Fig. 3 the variation in their COP (i.e. difference between max and min) under different loading regimes, and note that there is a large fluctuation even for a given load. For e.g., COP for chiller 1 varies between 5.7 and 8.2 for 70% load, and between 1.8 to 8.3 for 60% load. This is because the chiller COP in practice is highly dependent on a variety of factors. We also observe from the data that if we classify the COPs at 5% cooling load increments, then more than 40% of the COPs are missing.

2.3 Possible Benefit of Accurate COP Prediction

Chillers are complex systems. The COP fluctuation observed above is the result of thermodynamic processes under changing environmental conditions and configurations of the local building context, as well as the impact of other parameters such as chiller degradation and exposure to different seasons/weather. These factors are exceedingly difficult to capture within an analytical model. Data-driven techniques can thus play a crucial role in accurate COP prediction for improved chiller sequencing in the real world.

TABLE 2
Example of a 10 × 5 updated profile on 2015.12.31 in Pacific Place II, with out-of-date entries discarded.

Load	Chiller 1	Chiller 2	Chiller 3	Chiller 4	Chiller 5
50%	7.4	-	-	7.6	7.0
55%	-	-	-	-	-
60%	6.4	7.1	4.7	6.9	-
65%	-	-	-	-	-
70%	7.3	5.6	7.2	-	-
75%	-	-	-	-	-
80%	-	6.9	5.4	-	-
85%	-	-	-	-	-
90%	-	-	7.0	-	6.6
95%	-	-	-	-	-

Currently, in practice, facility managers often perform COP estimation using initial profile. There are two issues with current estimation schemes: 1) not all loads have been tested in the initial profiling period, and so COPs at these loads will be missing. A consequence is that these loads will never be used in the sequencing algorithm; 2) for the COPs with data, a simple averaging approach is often used. A consequence is that these COPs may be largely inaccurate and should not be used for making sequencing decisions.

In this section, we demonstrate that there can be a substantial reduction in chiller electricity consumption when sequencing is performed with accurate COP prediction.

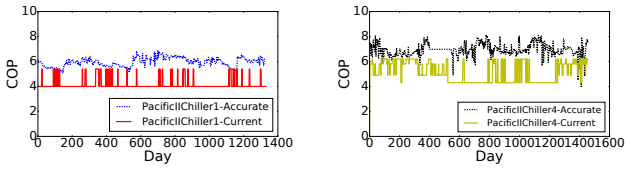
We now compute the electricity consumption using chiller sequencing under the current COP estimation scheme, which is based on the initial profiles, and compare it against a scheme that estimates COP accurately assuming there exists an oracle that can determine these values. Such an oracle can be obtained by computing COP using historical data (Section 2.1), and can be regarded as ground truth. Clearly, this is not a fair comparison, but it shows the benefit that comes with improving COP prediction.

Since the current chiller sequencing mechanism uses COPs that are inaccurate, it is possible that the cooling load Q provided by the chillers fails to satisfy the actual cooling demand $\mathcal{D}$. In practice, this is usually addressed by starting backup chillers immediately. [1]

[A case study is conducted based on Pacific Place II, which contains the most complete chiller data amongst the three buildings. A recent COP profile, e.g., on 2015.12.31 for Pacific Place II is shown in Table 2. We can see that the in-use COP profile for that day is indeed sparse, confirming that each chiller is routinely operated at only a few distinct loads.]

To conduct sequencing in Pacific Place II, COP estimation is needed for all the entries in the table. In Fig. 4, we start by comparing the predicted COP in the current mode of operation (bottom curve) against the accurate oracle scheme (top curve). The curves depict chiller 1 (left) and chiller 4 (right) in Pacific Place II. We see that the current mode of operation often under estimates the COP. There is a high estimation error of over 35% and has little correlation

1. It is possible to over-provision the cooling loads by a certain degree. There is a trade-off between over-provisioning and the fail-over for backup chillers. This problem is orthogonal to ours and we do not consider it in this paper.



(a) (b)

Fig. 4. The accurate COP and estimated COP with present scheme in Pacific Place II for two types of chillers (a) Chiller 1; (b) Chiller 4.

with the top curve. A perfect prediction could have resulted in lowering electricity consumption by over 45% on average, with nearly 60% reduction in the year 2015.

2.4 BT-DCS Problem Definition

1. The Data-driven COP Prediction (DPP) Sub-Problem. We have seen that an accurate estimation of the COP leads to substantial reduction in chiller electricity consumption and chiller sequencing heavily relies on the COP. Our idea is to develop individualized COP for each chiller by applying machine learning techniques using historical chiller data. A server is established to store the historical data from the BMS. When a cooling demand $\mathcal{D}$ arrives, the server can perform chiller sequencing assisted by our data-driven COP prediction schemes. To this end, we first introduce the Data-driven COP Prediction Sub-Problem.

Formal definition of DPP Sub-Problem: Given the current prediction task, infer the COP profile $\boldsymbol{COP}$ which minimizes the prediction loss, i.e., $1/T \sum_{t<T} ||F_t(\boldsymbol{X}_t, \boldsymbol{W}) - COP_t||_2$, where $\boldsymbol{X}$ denotes the features at time t of total time T; $\boldsymbol{COP}_t$ denotes the predicted COP for all chillers at time t; $F(\cdot)$ denotes the learning and prediction model and $\boldsymbol{W}$ denotes its parameters.

2. The BT-DCS Problem. The next step is to determine the optimal sequencing of chillers. In practice, one needs to be wary of prediction *deadline* T_D which is defined as the total time length of one chiller sequencing operation, including the computation time and the mechanical switching time, computed considering both the periodic interval t_P and mechanical switching time t_M, e.g., $T_D = \min(t_P, t_M)$ [15]. Moreover, we also consider the system budget P_B, which is defined as the total cost to maintain the prediction service, including the establishment fee of additional computing resources p_e, the data storage fee p_s and the data processing fee p_p, e.g., $P_B = p_e + p_s + p_p$

Formal definition of BT-DCS Problem: Given cooling demand $\mathcal{D}$, deadline T_D, budget P_B, historical COP labels $\boldsymbol{COP}$, the targeted and historical feature values $\boldsymbol{X}_c$, $\boldsymbol{X}$, our objective is to find a chiller sequence $\boldsymbol{Q} = \{Q_i\}$ which minimizes the total energy consumption E. The final solution should satisfy (1) the cooling demand, i.e., $\sum_i Q_i > \mathcal{D}$, (2) the total needed time T is within the deadline T_D, i.e., $T \leq T_D$, and (3) the system fee P is within the budget P_B, i.e., $P \leq P_B$.

In the following sections, we focus on (1) Data-driven COP Prediction for the DPP Sub-Problem (Section 3) and (2) Edge-based Data-driven Chiller Sequencing for the BT-DCS Problem (Section 4).

3 Data-driven COP Prediction for DPP

To solve the DPP Sub-Problem outlined in Section 2.4, we develop an approach involving two steps: *Domain-assisted Feature Engineering* and *Clustered Multi-task Learning*, as described next.

3.1 Domain-assisted Feature Engineering

In industry domain, there is usually no luxury to have enormous data where a model can be trained to automatically eliminate irrelevant features. As such the first challenge is to select the proper feature set for chiller performance profiling. Our feature engineering uses domain knowledge to create features relevant to the problem at hand. The understanding includes the influence of external environmental conditions and the influence of inner mechanical factors associated with the chillers. We list our features in Table 3.

Temporal Features. First, we exploit the seasonality and the age of the chillers (in terms of days) as the temporal features. Intuitively, the chiller demands exhibit distinctive temporal characteristics: 1) cooling loads of chillers are different in different seasons, especially summer and winter, which leads to varying performance degradation; 2) as chillers age, its performance gradually degrades as well [10].

Meteorological Features. Second, we know that meteorological information such as temperature and weather drive the cooling demand imposed on the chillers. For example, a higher outdoor temperature requires more cooling power to ensure a comfortable room. This meteorological factor would affect the chiller mode and thus the chiller performance.

Mechanical Features. Finally, mechanical features are used to capture the chiller characteristics. The model type, building, operating power, water temperature difference, flow rate and the recent cooling load are important features. The cooling load is the amount of heat energy that would need to be removed from a space to maintain the temperature within an acceptable range. [In practice, cooling loads are assigned to air-conditioning equipment, i.e., chillers, in order to meet the overall cooling demand, which reflects the amount of work that chillers provide, and thus significantly impacts chiller degradation.]

3.2 Clustered Multi-task Learning

Another challenge we face is the sparsity of the performance profile. Reliable sequencing requires the COP for all chiller loads. However, as shown in the Table 2, it is common for chillers to run on merely a small distinct set of loads, which leaves a sparse profile for training and prediction. The COP corresponding to the empty loads is difficult to infer reliably with little training data.

A natural way to solve this sparse problem is to infer the values using other non-empty ones. However, simply replacing with neighboring non-empty entries can cause significant errors. For example, in Table 2, we see that even for the same Chiller 3, replacing the COP of 80% with 90% leads to a relative error of 29.63%. That is because, at the time when these COPs are updated, external environment (e.g., meteorological factors) and inner conditions (e.g., temporal and mechanical factors) can be different, which leads to

TABLE 3
The description of features.

Feature Type	Feature	Description
Temporal	Season	The season which the time interval is in
	Age of chiller	The number of days that the chillers have been working
Meteorological	Weather Condition	The description of weather condition in a time interval
	Outdoor Temperature	The outdoor temperature measured by Celsius in a time interval
Mechanical	Model Type	The model of the operating chiller
	Building	The building that the operating chiller is deployed in
	Operating Power	The power measured by kilowatts for the operating chiller
	Water Mass Flow Rate	The mass of water flowing per second, measured by kg/s
	Water Temperature Difference	The difference between the returned and supplied chilled water temperature
	Latest Cooling Load	The last recored cooling load assigned on this chiller

different thermodynamic processes of operation, resulting in different COPs. In other words, for different entries, different model parameters must be used to capture the underlying thermodynamic processes, i.e., we need to train one model for each entry, while at the same time exploit the benefits that come with (potentially) more information being present in larger training data sets.

[Specifically, our idea here is to learn from not only the training data available for this single entry, but also learn from training data present in other pertinent contexts, e.g., cases with similar temporal, meteorological and mechanical conditions. To this end, we apply *Clustered Multi-Task Learning (CMTL)* approach [16].] The COP prediction on an entry is called a *prediction task* in our paper. For each entry, the prediction task collects training data from similar contexts. CMTL is suitable for our multi-task and sparse condition, i.e., we not only need to develop different model parameters for each entry, but also need to share knowledge, e.g., training samples, among these entries. Besides, such a learning approach can adopt any machine learning models, e.g., SVM or AdaBoost, as its base model. [2]

4 Edge based Data-driven Sequencing under Budget and Time constraints for BT-DCS

4.1 Cloud-based Solution is Expensive for BT-DCS

Above we have described a solution for the Data-driven COP Prediction (DPP) sub-problem. A natural thinking is using the cloud to deploy DPP and finally solve the BT-DCS problem. With the consideration of the time constraint T_D, a sequencing solution with time constraint based on dominant graph is proposed in [15], which helps to reduce the inference time by merely predicting a subset of all tasks. The Time-constrained Data-driven COP Prediction takes the historical data $\boldsymbol{X}$, historical COP labels $\boldsymbol{COP}$, deadline T_D as input and generates the COP prediction result within the deadline. Then, using the predicted COP and cooling demand $\mathcal{D}$, Sequencing Determination outputs the optimal chiller sequence and Feedback Control ensures the successful execution of the sequencing in the local HVAC plant.

However, to solve the BT-DCS, we also need to consider the budget constraint P_B. However, the above cloud-based time-aware solution may exceed the budget. It need to collect data sets from different buildings and also require the computation power of cloud to process the whole data set. More specifically, (1) In many cases where we have buildings managed by different BMS providers, it is expensive and even impossible to merge the data from all these different buildings together, leading to problems on scalability of the data-driven sequencing. (2) In our example, even for a single BMS provider in charge of 10 buildings, finishing dominant-graph-based T-DCS within the typical chiller sequencing period of two hours requires 29 m4.xlarge instances and 10-TB General Purpose SSD (gp2) volume on Amazon Web Service (AWS), for which the total annual price is $\$41,325$ (m4.xlarge) $+\$14,746$ (gp2)$=\ \$56,071$. Considering the two issues, the cloud-based service is still quite expensive and is likely to break the budget in BT-DCS.

4.2 Existence of Redundant Data

In this paper, we propose to solve the expensive problem by considering edge-based architecture. The idea behind is that we only conduct COP prediction on the edge by leveraging existing computing resources, which is much less expensive. The question is the amount of data set we should use, in order to maximize the overall optimality of our DCS approach under the time and budget constraint.

Fig. 5 shows the electricity consumption of data-driven sequencing with 35% and 70% coarsely discarded training data on the cloud. We see just 2.37% and 4.23% higher electricity consumption compared with the complete raw data, respectively. In the following, we will show that by carefully selecting data to train, the cost of cloud service for our approach can be significantly reduced.

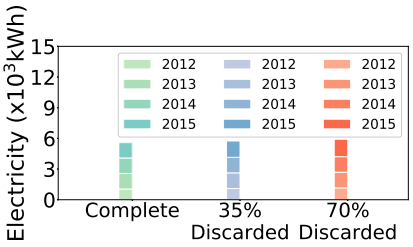


Fig. 5. Electricity consumption with 1) all raw data, 2) 35% coarsely discarded, and 3) 70% coarsely discarded.

The key observation of our approach is that, while in principle all data may be used to select the chiller sequencing, in practice only a small subset of them are frequently used in the data-driven sequencing. When sufficient data is collected, the accuracy of COP predictions only shows very marginal improvement and so the energy savings begins to plateau. Figure 6 shows the result.

We see that in general, the accuracy of a COP prediction task increases as the number of chillers, and in turn the data used for training, increases; but the improvement

2. For interested readers, the prediction results using different base models are reported in [15].

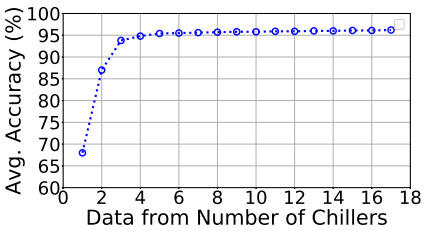


Fig. 6. The average prediction accuracy as a function of number of chillers data used.

tapers after data is collected from five chillers. The reasons behind this are (1) different buildings have different cooling demands and experience different environment conditions. Accordingly, not all chillers, with different makes and cooling capacities, render themselves useful for transfer learning. Worse still, even if the model types of chillers are the same, they operate under different configurations, which leads to varying thermodynamic models and negative transfer. (2) In a given building, the chillers are typically sourced from a single vendor, tend to be of the same generation and experience similar environmental conditions. As a result, data from these chillers provide important training samples that become valuable for knowledge transfer.

Intuitively, one should *reduce* the redundant data used in the system, which then lowers the high service cost including establishment, usage and maintenance fee of cloud-based transmission and storage. We have introduced a dominant-graph approach to rank and focus on the reduced entries in the COP profile in Table 2, while still needing to merge and process the databases from different buildings. In this study, for data-driven chiller sequencing, we aim to develop an inexpensive architecture on the edge to reduce the service cost by making better use of valuable edge data, instead of transmitting, storing and processing the whole raw (and mostly redundant) database on the cloud.

4.3 Edge-based DCS (E-DCS) Framework

Based on the above observations, we propose the Edge-based DCS (E-DCS) approach, framework for which is shown in Fig. 7. In this framework, there are four parties: the cloud, the edge node (E-node), the network node (N-node), and the sensing node (S-node).

The Cloud: Many cloud service providers can rent the cloud computing and storage resources in a pay-as-you-go model, like AWS, Alibaba, etc. The key module of the cloud is the network Operations, Administration, and Management (OAM) module. It computes the network topology, that is, the peering thing-to-cloud communication (TCC) links between itself and multiple E-nodes. For these TCC links, there are network protocols such as IEEE 802 families (like Ethernet, Wifi), BAC-net, LTE Category 1 (CAT1), LTE Category 0 (CAT0), or other standards specified by the Generation Partnership Project (3GPP) that we can use. The data flow from multiple E-nodes are stored in database. The advanced COP prediction module performs prediction tasks based on the multiple building's data which can be used as an optional choice depending on the architecture. Last, the TCC module maintains the data link level connection between itself and multiple E-nodes within its communication range.

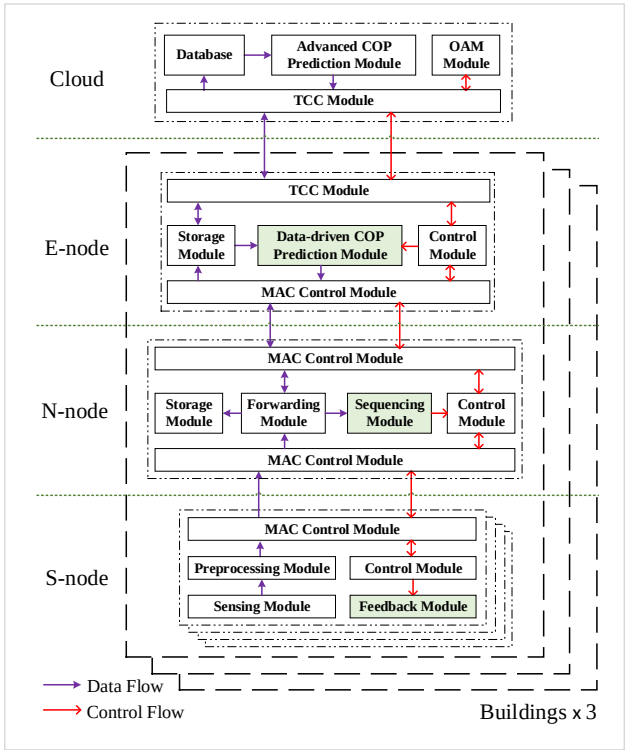


Fig. 7. Framework of Edge-based Data-driven Chiller Sequencing.

The E-Node: In our case, the edge node can be located and serve in the private data center for the BMS service provider. There are mainly five modules. The TCC module maintains the data link level connection between itself and the cloud. The MAC control module maintains the data link level connection between itself and N-nodes. The control module answers network layer queries from N-nodes. Another is the storage module which is used to store the data from the cloud and N-node. Last, the data-driven COP prediction module performs the machine learning prediction tasks and returns the result to the N-node for sequencing determination.

The N-Node: The Network node can be a router or gateway in each building. There are mainly five modules. The MAC control module maintains the link-level connections among the E-nodes, the S-nodes and itself. These LOC links can employ WiFi, Bluetooth, Zigbee, and so on. The forwarding module mainly forwards the data from S-nodes to E-nodes for prediction and to storage module for storing. In addition, it also forwards the prediction result from E-node to sequencing module for performing the sequencing algorithm to select the optimal sequencing. Then, the sequencing module sends the optimal sequencing instruction to the S-node through the control module. Last, the control module answers the network layer queries from S-nodes.

The S-Node: The sensing node is located on each equipment, e.g., chiller and pump. It collects sensing data, then it sends those sensing data to the preprocessing module for data cleaning and could be used as the input to various machine learning models. The control module maintains the network topology and issues sequencing instructions to the feedback module. Last, the feedback module receives the instruction and adjusts the power of each equipment.

4.4 Design and Discussion on Architectures: Instances of the Edge-based Framework

For different scenarios, we provide different edge-based architectures and the corresponding discussion on them.

1. Cross-building Architecture. First, we introduce the architecture of cross multiple buildings, shown in Fig. 7. In this architecture, data from different buildings are leverage to predict the COP and a unified model is trained to perform predictions for all loads of chillers.

This architecture provides the data sets from different buildings, which can theoretically support machine learning techniques using cross-building or single-building data. A key discussion is then the allocation of the prediction module. (1) When the prediction module is allocated on the cloud, the multiple-buildings data will be uploaded to the cloud. Such a centralized solution collects and processes all data on the cloud, which leads to high cost of cloud service but provides the highest possible accuracy and energy-saving. (2) When the prediction module is allocated on the edge, the other multiple buildings data will transfer to the corresponding edge through the cloud. Such a design only requires cloud to provide data storage service which costs much less. But it can lead to high overhead of cross data transmission between each possible pairs of buildings and the cloud. This architecture works well especially when we need to merge data from different buildings, e.g., data of a certain type of chiller is not sufficient in a single building, and the related fee on the cloud is not expensive.

2. Single-building Architectures. Another design is to conduct data-driven prediction and sequencing merely within a single building, which usually includes the following two variant architectures, namely single-chiller architecture and cross-chiller architecture, as shown in Fig. 8 and Fig. 9.

2.1. Single-chiller Architecture. The main idea of this type of architecture is to conduct data processing and analysis closely to the source of data generation so as to reduce redundant data transmission and storage. It includes two variant architectures, namely dynamic-model architecture and fix-model architecture, as shown in Fig. 8.

(a) Dynamic-model Architecture. As shown in Fig. 8.(a), we put the prediction module closely to the source of data generation, e.g., S-nodes. We applied the transfer learning technique to dynamically capture different models on each single S-node, the result of which is summarized and transmit to the N-node for sequencing determination.

Such an architecture helps to reduce the amount of data needed for learning as much as possible. However, the prediction performance may not be good enough due to the incomplete data profile in each chiller. While transfer/online learning methods are used to overcome such a deficiency, it also requires higher computation complexity and data storage.

(b) Fix-model Architecture. The main idea of this architecture is to off-line training a fixed model for each single S-node using unified machine learning techniques which is indicated as Fig. 8.(b). It then sends the prediction results to the N-node for sequencing determination.

This architecture does not need data to change its model and can even have models pre-installed in the S-node. It is thus possible to get rid of data storage and transmission and run in the lowest price. However, when the data are incomplete, the results of the fixed model is not reliable.

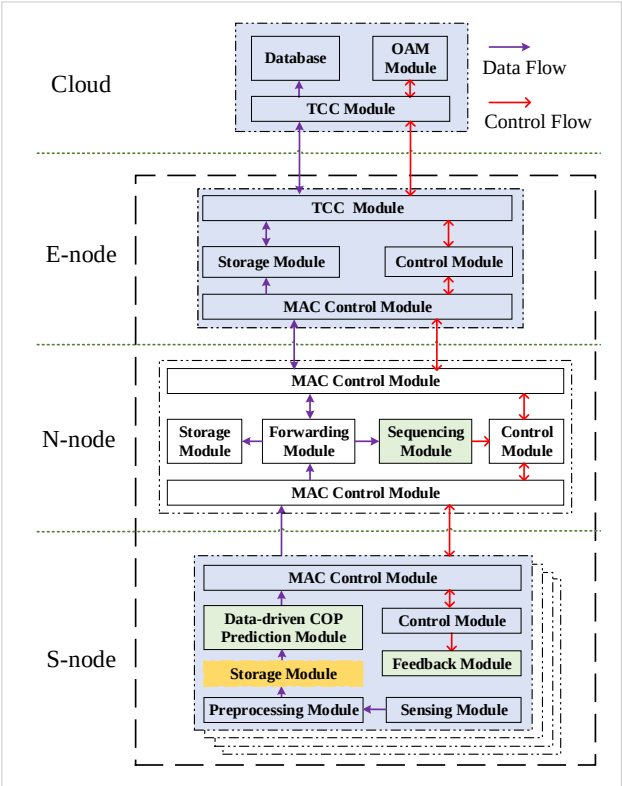


Fig. 8. Two Variants of Edge-based Single-building Single-chiller Architectures for DCS: (a) Dynamic-model Architecture with Storage Module; (b) Fix-model Architecture without Storage Module. Differences against the framework are marked in blue.

2.2 Cross-chiller Architecture. Usually, the sensing data of single chiller is not sufficient for learning, e.g., as shown in Table 2 in a previous section. In this architecture, all data in a single building are trained in the edge server instead of on the cloud, as shown in Fig. 9, which therefore reduces the system fee as much as possible.

The architecture can theoretically support machine learning techniques using cross-chiller data. When sharing similar sensing data across multiple chillers using CMTL, it can overcome the deficiency of data scarcity issue and improve the prediction accuracy. It is particularly useful when there are sufficient data in a single building, e.g., the building have multiple chillers of the same type and the learning process is not expensive or even free within the building. Based on our example, we select this architecture which maximizes the optimality of the BT-DCS.

5 Performance Evaluation

Experimental Setting. The total data collected from the BMS is more than 1 TB. We configure a private cloud to process the data for our experiments, with 16 cores of 2.6GHz CPU and a total memory of 64GB. We train the models with three-year data and predict with one-year data, which is a common setting in time-series data mining [17] and multi-task learning [18].

Baselines. To make the COP prediction, we employ the following state-of-the-art models as baselines.

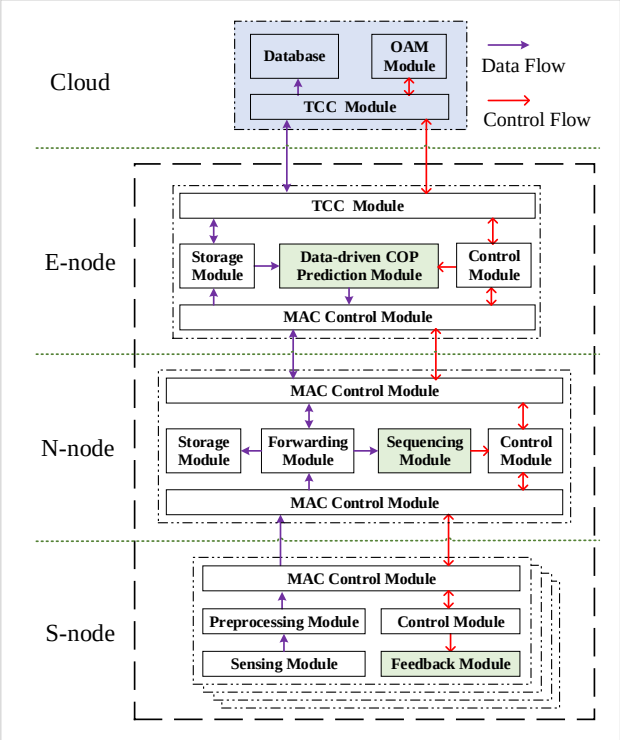


Fig. 9. Edge-based Single-building Cross-chiller Architecture of Data-driven Chiller Sequencing. Differences against the framework are marked in blue.

- **Initial Profiling Model (IPM)** predicts the COP using the initial profile of chillers under different loads.
- **Thermodynamic Model (TDM)** predicts the COP using pre-calibrated thermodynamic model. Thermodynamic models [9] capture the thermodynamic process of chillers and try to obtain the chiller COP with fixed form given the chiller loading.
- **Data-driven COP Prediction Model on Cloud (DPP)** predicts the COP by the data-driven approach which learns the model with historical data samples, where **DPP-Ada**, **DPP-SVR** denotes the approach using AdaBoost, SVM Regression as learning model, respectively. This approach is running on the cloud.
- **Time-constrained DPP Model on Cloud (T-DPP)** predicts the COP by data-driven approach under time constraints. As for the default approach, we adopt the Joint Priority Ordering method to order entries [15]. This approach is running on the cloud.
- **Edge-based DPP Model (E-DPP)** predicts the COP by data-driven approach on the edge with consideration of time and budget constraints. As for the default approach, we deploy the prediction and sequencing based on our proposed Cross-chiller Architecture.

To leverage the estimated COP, we employ the following state-of-the-art sequencing models as baselines.

1) **Predefined Sequencing (PS)** conducts sequencing with predefined prediction model and starts backup chillers when it fails to meet the cooling demand [19], [20]. We adopt Thermodynamic Model as default predefined prediction model instead of Initial Profiling Model because it performs better.
2) **Data-driven Chiller Sequencing on Cloud (DCS)** conducts sequencing with DPP and predicts all profiles without considering any time-constraint. We adopt the most accurate DPP-Ada as default DPP model and use backup chiller for sequencing. This approach is running on the cloud.
3) **Time-constrained DCS on Cloud (T-DCS)** conducts chiller sequencing with DPP under time constraint. We adopt the Joint Priority Ordering method to select entries and start backup chillers when necessary [15]. This approach is running on the cloud.
4) **Edge-based Data-driven Chiller Sequencing (E-DCS)** conducts chiller sequencing with DPP under time and budget constraint. We adopt our proposed Cross-chiller Architecture and start backup chillers when necessary.

Evaluation Metrics. For a sequencing method, the ability to provide credible energy saving is crucial to all stakeholders. Electricity is always the first concern, and we measure the Average Electricity Consumption (AvgEC), which measures the average electricity used by all sequencing operations on one day where all time instances in one day is denoted by T and each sequencing is conducted at time $t \in T$. Let L_i denote the maximum cooling capacity of chiller $i < n$. $COP_{i,t}$ denotes the real performance of chiller i at time t. Formally,

$$AvgEC = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} L_i \cdot S_{i,t}/COP_{i,t}.$$

It is also significant that our decision should be accurate so that our final decision making can be reliable. Thus, we also measure the Accuracy, which indicates the similarity between our predicted COP and the real COP. $\hat{COP}_{i,t}$ denotes its predicted value. Formally,

$$Accuracy = \frac{1}{T}\sum_{i<n}\frac{|COP_{i,t} - \hat{COP}_{i,t}|}{COP_{i,t}}.$$

Our decision should be conducted before the deadline, and we also measure the two metrics on time: Total Time and Run Time. Total Time indicates the total time over which sequencing was made, including the computation time and the mechanical switching time, to see whether the proposed operation can be done within time limitations. Run Time indicates the computation time of prediction models, and thus indicates the power of searching methods, to see whether all of the prediction tasks should be done. Formally,

$$Total\ Time = t_s - t_c, \quad Run\ Time = t_p - t_c,$$

where t_s denotes the time instant when the sequencing decision is made; t_p denotes the time instant when the predicted COP is known; t_c denotes the time when each experiment starts.

Last, our approach should be appropriate for actual deployment and reduce the system fee as much as possible. Thus, we also compare the System Fee, which indicates whether the approach we adopted is more cost-effective. Formally,

$$System\ Fee = p_e + p_s + p_p,$$

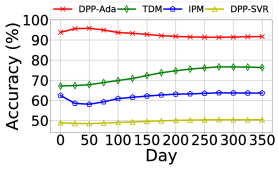


Fig. 10. The accuracy as a function of day.

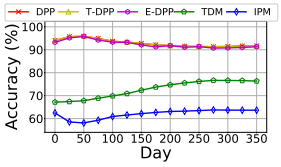


Fig. 11. The accuracy of prediction models as a function of day.

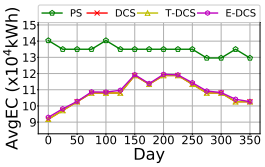


Fig. 12. The average electricity consumption of days.

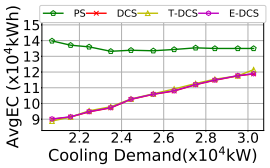


Fig. 13. The average electricity consumption of cooling demands.

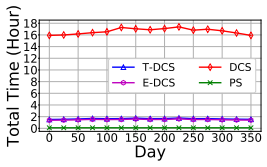


Fig. 14. The total time as a function of day.

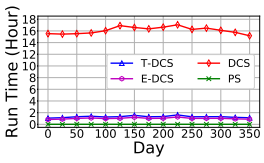


Fig. 15. The run time as a function of day.

where p_e denotes the establish fee of additional computing resources; p_s denotes the storage fee; p_p denotes the data processing fee.

5.1 DPP and E-DPP Model

We first compare the prediction results of our Data-driven COP Prediction (DPP-Ada and DPP-SVR) with that of Initial Profiling Model (IPM) and Thermodynamic Model (TDM), and Edge-based DPP will be evaluated next. Figure 10 shows that, DPP-Ada outperforms DPP-SVR, Thermodynamic Model, and Initial Profiling Model by 43.19%, 20.14%, and 30.77% respectively on average, which illustrates the prediction power of our approach. That is because, our data-driven approach is developed based on the runtime data in the real environment and leverages the ensemble technique to avoid over-fitting in non-linear modeling, which can successfully capture the chiller local and dynamic performance.

In Fig. 11, we compare the Accuracy of Edge-based DPP (E-DPP) with that of DPP (we adopt DPP-Ada as our default DPP approach due to its high accuracy), T-DPP, Initial Profiling Model and Thermodynamic Model for one year. Though E-DPP significantly reduces the computation time (will be shown later), we can see that our E-DPP is comparable with DPP and T-DPP in terms of accuracy, which outperforms Initial Profiling Model and Thermodynamic Model by 31.98% and 20.72%. That is because E-DPP leverages the state-of-the-art data-driven model, which maintains the superiority of our data-driven techniques and it meets the time and budget constraints by selecting necessary tasks and data to process which does not sacrifice the accuracy.

5.2 DCS and E-DCS Model

Result on Electricity Consumption. Along with different days of time, Figure 12 shows the changes in Average Electricity Consumption with different sequencing methods. First, on average, we see that our DCS outperforms Predefined Sequencing (PS) by 32.04%. E.g., in the 25th day, the improvement increases to 38.89%. That is because our data-driven method captures the performance dynamics of chillers and adjusts the cooling load in a smarter way. The Predefined Sequencing remains stable because its backup chiller mechanism is triggered frequently due to low prediction accuracy of COP and thus consistently provides more cooling than needed. Second, though our Edge-based DCS (E-DCS) significantly lowers system fee and reduces the computation time (will be shown later), we see that it remains comparable with DCS and T-DCS in terms of electricity usage, which always outperforms Predefined Sequencing, because E-DCS also maintains the superiority of data-driven techniques as DCS methods.

Under changing cooling demands, Figure 13 shows the changes in Average Electricity Consumption with various sequencing methods. We see that E-DCS is still comparable with DCS and T-DCS, and always outperforms Predefined Sequencing, which confirms the performance of our approach mentioned above. For example, E-DCS outperforms Predefined Sequencing by 23.67% at the average cooling demand of 26205.5 kW. As the cooling demand increases, the Average Electricity Consumption of E-DCS, DCS and T-DCS also gradually increases; while Predefined Sequencing still remains steadily high due to its limited prediction performance and backup chiller mechanism.

Result on Time. We first compare the total time of the-state-of-the-art sequencing models. In Fig. 14, we compare the Total Time of E-DCS with that of Predefined Sequencing, T-DCS and DCS for one year under time limitation of 2 hours. We can see that E-DCS, T-DCS and Predefined Sequencing operations can be completed within the stipulated time, but DCS can not. It is mainly because DCS needs to update all the profile entries before making a sequencing decision, while our E-DCS not only updates mere the most important profile entries, but also reduces the amount of data for model training, which saves a significant amount of computation time.

To show the potential of saving time, we next compare the Run Time of E-DCS with that of Predefined Sequencing, T-DCS and DCS for one year. As we can see in Fig. 15, the average computation time of our E-DCS model is 1.2 hours, which is an improvement of 14 times over the DCS model. That is because, E-DCS not only uses Joint Priority Ordering to select the most important entries for prediction, but also reduces the amount of data for model training in each period, unlike DCS. Though compared with Predefined Sequencing, our E-DCS takes more time, but in return, we get a more accurate result as detailed in Section 5.1.

Result on System Fee. Figure 16 shows Annual System Fee of E-DCS (Edge-based architecture) and T-DCS (Cloud-based architecture). As we can see, though E-DCS may cost more on the establish fee of additional computing resources, e.g., a general edge server, the data storage and processing fee of E-DCS is significantly reduced compared with T-DCS. Specifically, the data processing fee of E-DCS can be quite close to 0 since it put all the computation on the edge as much as possible. All in all, combined with the price of AWS, E-DCS saves 43.67% Annual System Fee compared with T-DCS in our estimation. These results highlight the

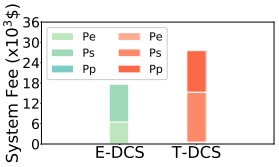


Fig. 16. The average annual system fee comparing E-DCS with T-DCS.

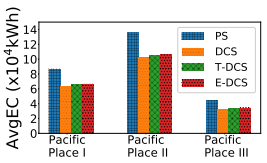


Fig. 17. The average electricity consumption in multiple buildings.

cost-effective of our E-DCS approach. When it comes the building have multiple chillers of the same type, our E-DCS approach achieves the highest accuracy and energy saving using CMTL on multiple chillers, while it meets the time constraint using dominant-graph approach; it does not have to merge the data under different buildings and does not need data to train on the cloud, which therefore significantly reduces the system fee to pay each year.

Result on Multiple Buildings. In Fig. 17, we compare the Average Electricity Consumption of E-DCS with that of Predefined Sequencing, DCS and T-DCS for different buildings. As we can see, though E-DCS significantly reduces the computation time, the Electricity Consumption of E-DCS is still quite close to T-DCS and DCS. Compared with Predefined Sequencing, E-DCS saves 20450 kWh of Electricity Consumption in Pacific Place I, which is an improvement of 30.48%, respectively. In the remaining two buildings of Pacific Place II and Pacific Place III, the improvement is 30.68% and 30.12%, respectively. These results highlight the generality of our approach. When it comes to multiple buildings, our approach merely shares the similar training samples using multi-task learning and selects important entries under similar cooling demands, thus avoiding the noise when switching between different contexts and models.

6 Related Work

Chiller Sequencing refers to operating the most efficient combination of chillers in a building at (near) real-time to meet the time-varying cooling demand. Previous studies mainly focused on developing reliable and robust sequencing according to instantaneous building cooling load [21], [22], [23]. These studies mainly set the chiller cooling capacity as constant (i.e., the same as the rated cooling capacity) [24]. Considering that the cooling capacity may vary under different operating conditions (e.g., different chiller evaporating and condensing temperatures), such approaches may fail to provide enough cooling energy or lead to extra energy usage [25]. As a solution, first, thermal energy storage was leveraged to improve the COP of multiple chiller plants [26]. [Second, physical and grey box have been used to capture the variation in maximum cooling capacity given different operating conditions [25], [27].] General model is also proposed calibrated using real data from water plants [28], [29], [30] and centrifugal chillers [31]. However, the actual performance of these sequence control strategies is subject to the accuracy of these models because they are general purpose models and do not capture the practicalities that come with deployments in different building conditions and time-varying cooling demands. Worse still, when conducting data-driven sequencing, they also do not consider the time limitations, e.g., minimum start-stop-start time of chillers. For the first time, we introduce a novel data-driven chiller sequencing framework that also captures the need to perform chiller sequencing under time constraints. It provides dynamic cooling performance estimation given a set of possible cooling loads, other configurations and environmental settings, and conducts sequencing under practical time limitations.

Edge Computing. Driven by the development of the Internet of Things, edge computing has becomes a new computing paradigm which blows up the networking community and becomes a heated research topic. In existing works, there are different research focuses for edge computing, e.g., finding a proper trade-off between the energy consumption and the execution delay [32], minimizing the overall application execution cost [33], distributed machine learning systems [34] and etc.

To determine the optimal sequencing of chillers on the edge, one also needs to be wary of the time and cost issue: (1) In general, the balance between time and cost is always an important problem for decision making in traditional complex system research [35], [36], and now need to be re-thinked, especially when time-costly machine learning introduced on the edge of the network [37]. (2) The cooling demand changes over time, so chiller sequencing must be performed repeatedly in order to continuously meet the varying cooling demand. The common practice is to trigger chiller sequencing in a periodic manner [21]. (3) To ensure cooling performance, chiller sequencing needs time for feedback control until the system regains stability when switching from one sequence to another. There is also a minimum start-stop-start time (called *deadband*) for every chiller. (4) The chiller sequencing for each period must be completed before the start of the next sequencing period. Otherwise, the system can be unstable and return inaccurate data which can be detrimental to subsequent COP prediction and sequencing operations, as well as for the overall performance of the chillers.

Our idea is based on the observation that not all data are valuable for prediction. Focusing on electricity minimization under the time and budget constraints, we show that it is possible to lower the computation fee by reducing the amount of less-important data and tasks, rather than sacrificing prediction accuracy or incurring high computation cost, which sheds some new light on machine learning in decision making on the edge.

7 Discussion

Accuracy and Energy Consumption. Prediction accuracy generally affects energy consumption in our BT-DCS sequencing, because additional operations, e.g., backup chillers, should be launched to fix the problem to meet the required cooling demand. Theoretically, it seems that using backup chiller does not necessarily lead to increase in overall energy consumption, when they are used to satisfy the exact cooling demand. However, in practice, starting additional backup chillers usually increases the energy consumption under perfect prediction due to over-provisioning convention, the launching and maintaining cost of backup chillers. However, there are also cases when inaccurate prediction may not lead to inefficient operation. For example,

some operations may never be used in chiller sequencing optimization. The prediction accuracy of such unimportant operations may not affect the final decision. It would be an interesting future work to further investigate on the difference of data-driven techniques between industrial (focusing on industrial objective) and traditional (focus on merely accuracy objective) prediction.

Chiller Type. In this paper, we focus on merely water-cooled chillers. [When conducting data-driven COP prediction, there are quite a few issues to consider for such a widely used type of chillers. First, we take into account the inlet and outlet water temperatures, which are two factors of water-cooled chillers directly affecting the COP. Second, we also leverage the dry-bulb temperature (DBT) to compute cooling demands. The varying DBT directly indicates the change of the outdoor environment and the cooling demand, so that can be used for performance prediction on the chillers. Third, we also consider the water temperature entering the condenser. It can be related to the COP of a water-cooled chiller. For example, a lower entering temperature usually leads to a higher COP, because it facilitates better heat emission and thus more effective refrigerating cycling. However, such a temperature is usually under control as a fixed value in a modern HVAC system. Take condensers in our case as an example, the entering temperature is set at around 32 degrees Celsius and the leaving temperature is set at around 37 degrees Celsius. For each of the temperature, the difference over time is too small, i.e., within half a degree. As we have known, when detecting varying COP over time, leveraging a feature with a fixed value in a data-driven model does not affect the output for our model. To this end, we choose not to include the condenser water temperature in our model.]

[Future works can include the feature design for different types of chillers, e.g., features for air-cooled chillers. A possibly reused feature of air-cooled chillers is the DBT, which can also be quite related due to the heat transfer between air-cooled chillers and the outdoor air.]

Sensor Missing. The sensing data required by the COP computation are all accessible in our system. However, such an assumption may not be true in other systems where some of the required sensors, e.g., mass flow rates sensors, may not be not available. In that case, cooling load can be estimated indirectly by inverse physical models based on the power consumption of chiller motors, see [38].

[**Data Availability and Quality.** Admittedly, there exist a challenging issue that using several-year data for training suffers from data availability and quality. In urban computing, several methods are proposed to tackle the sensing problem, e.g., building more sensing infrastructures or leveraging crowdsourcing. But such approaches still suffer from high cost, privacy concern and even no guarantee of data quality and continuity. Our work leverages transfer learning to reduce training data to 1-year, but the sensing issue still exists. Future works can include using an initial profile and thermal dynamic model for life-long learning.]

[**Learning and Prediction Frequency.** In this paper, the data-driven prediction is assumed to be conducted in an online manner due to the sparse COP profile. More specifically, the data samples of all operations are likely to be sparse, the prediction model should be updated frequently to adapt to the accumulated samples under demands, configurations, and degradation over time. To maintain the high performance of our model, in this paper, we update the parameters each time before a chiller sequencing operation. It would be interesting future work to investigate the optimal learning and prediction frequency in data-driven industrial operations.]

Multiple-time Chiller Sequencing. In this paper, we focused on single-time chiller sequencing, i.e., without the consideration and cooperation of other sequencing at different times. Possible future work can include the consideration of minimum annual run-time of chillers for multiple-time sequencing, or minimum chiller utility for each single-time sequencing operation.

8 Conclusions

Developing energy efficient buildings has long been an important research topic as facility managers grapple with the problem of reducing their building's electricity bills. In this paper, we focused on one of the core problems in building operation, namely HVAC chiller sequencing, and made the following contributions.

First, we demonstrated that using chiller COP values from initial profiles can be detrimental from the point of view of HVAC electricity consumption. We subsequently stressed the need for quantifying COP at run-time.

Second, we showed that predicting COP accurately is a challenging problem, requiring considerable computation time and hardware resources. To provide a practical solution, we propose a data-driven COP prediction model along with an edge based chiller sequencing control framework under time and budget constraints, which opens the doors for HVAC electricity consumption reduction while enabling ease of use of the scheme for real-world deployment.

Finally, we evaluated the performance of our solution by applying it to BMS data, spanning 4 years, obtained from multiple chillers across 3 large commercial buildings in Hong Kong. We showed that our solution can save over 30% of HVAC electricity consumption compared to the current mode of chiller operation in the buildings. We believe that sequencing chillers using a data-driven approach for COP prediction offers a simple and effective mechanism for reducing the electricity consumption associated with operating the HVAC in large commercial buildings.

Acknowledgments

The authors would like to thank Dr. Abraham Lam and Adrian from Mega automation Ltd. for their valuable discussion and feedback. Parts of the work were supported by the Hong Kong Polytechnic University under Grant No.: 1-BBYX, and by the School of Computer Science and Technology, HUST, with the National Key Research & Development (R&D) Plan under Grant No.: 2017YFB1001703, and the National Natural Science Foundation of China under Grant No.: 61761136014 and 61722206, and by the Technical Innovation Department, Cloud BU, Huawei Technologies Co.Ltd. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Heating, Ventilation and Air Conditioning. http://eex.gov.au/technologies/heating-ventilation-and-air-conditioning/, 2018.

[2] Luis Pérez-Lombard, José Ortiz, and Christine Pout. A review on buildings energy consumption information. *Energy and buildings*, 40(3):394–398, 2008.

[3] Property Council of Australia. Survey of Operating Costs: Office Buildings. https://goo.gl/7oT1W7, 2018.

[4] Wilhelm Kleiminger et al. Smart heating control with occupancy prediction: How much can one save? In *Proc. ACM UbiComp'14*, pages 947–954.

[5] Arun Vishwanath et al. A data driven pre-cooling framework for energy cost optimization in commercial buildings. In *Proc. ACM e-Energy '17*, pages 157–167.

[6] Nair et al. Battery energy storage systems: Assessment for small-scale renewable energy integration. *Energy and Buildings*, 42(11):2124–2130, 2010.

[7] Zhaohui Liu, Hongwei Tan, Duo Luo, Guobao Yu, Jin Li, and Zhenyu Li. Optimal chiller sequencing control in an office building considering the variation of chiller maximum cooling capacity. *Energy and Buildings*, 140:430–442, 2017.

[8] Wikipedia. Coefficient of performance. https://en.wikipedia.org/wiki/Coefficient_of_performance, 2018.

[9] Kody M Powell, Wesley J Cole, et al. Optimal chiller loading in a district cooling system with thermal energy storage. *Energy*, 50:445–453, 2013.

[10] Nofirman Firdaus et al. Chiller: Performance deterioration and maintenance. *Energy Engineering*, 113(4):55–80, 2016.

[11] FW Yu et al. Optimization of water-cooled chiller system with load-based speed control. *Applied Energy*, 85(10):931–950, 2008.

[12] Yundan Liao, Yongjun Sun, and Gongsheng Huang. Robustness analysis of chiller sequencing control. *Energy Conversion and Management*, 103:180–190, 2015.

[13] Thomas Hartman. All-variable speed centrifugal chiller plants. *ASHRAE Journal*, 56(6):68–79, 2014.

[14] A Michopoulos et al. Three-years operation experience of a ground source heat pump system in northern greece. *Energy and Buildings*, 39(3):328–334, 2007.

[15] Zimu Zheng, Qiong Chen, Cheng Fan, Nan Guan, Arun Vishwanath, Dan Wang, and Fangming Liu. Data driven chiller sequencing for reducing hvac electricity consumption in commercial buildings. In *Proceedings of the Ninth International Conference on Future Energy Systems*, pages 236–248. ACM, 2018.

[16] Laurent Jacob, Jean-philippe Vert, and Francis R Bach. Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems (NIPS)*, pages 745–752, 2009.

[17] Yongxin Tong, Yuqiang Chen, Zimu Zhou, Lei Chen, Jie Wang, Qiang Yang, Jieping Ye, et al. The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. In *Proc. ACM SIGKDD'17*, pages 1653–1662.

[18] Jaime Carbonell Keerthiram Murugesan. Self-paced multitask learning with shared knowledge. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2522–2528, 2017.

[19] FW Yu and KT Chan. Economic benefits of optimal control for water-cooled chiller systems serving hotels in a subtropical climate. *Energy and Buildings*, 42(2):203–209, 2010.

[20] Jijun Zhou, Guanghua Wei, W Dan Turner, Song Deng, David E Claridge, and Oscar Contreras. Control optimization for a chilled water thermal storage system under a complicated time-of-use electricity rate schedule. 2005.

[21] Yongjun Sun, Shengwei Wang, and Fu Xiao. In situ performance comparison and evaluation of three chiller sequencing control strategies in a super high-rise building. *Energy and buildings*, 61:333–343, 2013.

[22] Yundan Liao et al. Uncertainty analysis for chiller sequencing control. *Energy and Buildings*, 85:187–198, 2014.

[23] Sen Huang, Wangda Zuo, and Michael D Sohn. Amelioration of the cooling load based chiller sequencing control. *Applied Energy*, 168:204–215, 2016.

[24] Pe Marchall Seymore. Simplified chiller sequencing for a primary/secondary variable chilled water flow system. *ASHRAE Journal*, 56:24–32, 2014.

[25] Yongjun Sun et al. Chiller sequencing control with enhanced robustness for energy efficient operation. *Energy and buildings*, 41(11):1246–1255, 2009.

[26] Madhur Behl, Truong X Nghiem, et al. Green scheduling for energy-efficient operation of multiple chiller plants. In *IEEE RTSS'12*, pages 195–204.

[27] Wei Jiang et al. General methodology combining engineering optimization of primary hvac&r plants with decision analysis methods - part ii: Uncertainty and decision analysis. *HVAC&R Research*, 13(1):119–140, 2007.

[28] Mark Hydeman, Nick Webb, Priya Sreedharan, et al. Development and testing of a reformulated regression-based electric chiller model/discussion. *ASHRAE Transactions*, 108:1118, 2002.

[29] Mark Hydeman et al. Tools and techniques to calibrate electric chiller component models/discussion. *ASHRAE transactions*, 108:733, 2002.

[30] Danielle Monfet and Radu Zmeureanu. Identification of the electric chiller model for the energyplus program using monitored data in an existing cooling plant. In *Proceedings of the international IBPSA conference. Sidney, Australia: International Building Performance Simulation Association*, 2011.

[31] JM Gordon et al. Centrifugal chillers: thermodynamic modelling and a diagnostic case study. *International Journal of refrigeration*, 18(4):253–257, 1995.

[32] Yuyi Mao, Jun Zhang, SH Song, et al. Power-delay tradeoff in multi-user mobile-edge computing systems. In *IEEE GLOBECOM*, 2016.

[33] Sowndarya Sundar and Ben Liang. Offloading dependent tasks with communication delay and deadline constraint. In *IEEE INFOCOM*, 2018.

[34] Kevin Hsieh, Aaron Harlap, Nandita Vijaykumar, et al. Gaia: Geo-distributed machine learning approaching lan speeds. In *NSDI*, 2017.

[35] Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. In *Advances in Neural Information Processing Systems*, pages 19–27, 2014.

[36] Michael Kearns et al. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[37] Abbas Kiani and Nirwan Ansari. Toward hierarchical mobile edge computing: An auction-based profit maximization approach. *IEEE Internet of Things Journal*, 4(6):2082–2091, 2017.

[38] Gongsheng Huang, Shengwei Wang, Fu Xiao, et al. A data fusion scheme for building automation systems of building central chilling plants. *Automation in Construction*, 18(3):302–309, 2009.

Zimu Zheng is currently a PhD student at the department of computing, the Hong Kong Polytechnic University. He received his B.Eng. degree in South China University of Technology, Guangzhou, China. His research interest lies in applied machine learning, e.g., edge AI and transfer learning, with an emphasis on IoT Data. He received the Best Paper Award of ACM International Conference on Future Energy Systems (ACM e-Energy) in 2018 and the Best Paper Award of ACM International Conference on Systems for Energy-Efficient Built Environments (ACM BuildSys) in 2018.

Qiong Chen received his B.Eng. degree in School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. He is currently a M.Eng. student in School of Computer Science and Technology, Huazhong University of Science and Technology. His research interests include applied machine learning and edge computing. He received the Best Paper Award of ACM International Conference on Future Energy Systems (ACM e-Energy) in 2018.

Cheng Fan received the B.Eng and Ph.D degrees in 2011 and 2016 respectively from the Hong Kong Polytechnic University, Hong Kong, China. Dr. Fan is currently an assistant professor in the College of Civil Engineering, Shenzhen University, Shenzhen, China. His research expertise includes building energy management, HVAC system controls and optimizations, and big data analytics.

Nan Guan is currently an assistant professor at the Department of Computing, the Hong Kong Polytechnic University. Dr. Guan received his BE and MS from Northeastern University, China in 2003 and 2006 respectively, and a PhD from Uppsala University, Sweden in 2013. Before joining PolyU in 2015, he worked as a faculty member in Northeastern University, China. His research interests include real-time embedded systems and cyber-physical systems. He received the EDAA Outstanding Dissertation Award in 2014, the Best Paper Award of IEEE Real-time Systems Symposium (RTSS) in 2009, the Best Paper Award of Conference on Design Automation and Test in Europe (DATE) in 2013 and the Best Paper Award of ACM International Conference on Future Energy Systems (ACM e-Energy) in 2018.

Arun Vishwanath (SM'15, M'11) is a lead research scientist at IBM Research in Melbourne, Australia working in the area of IoT for energy optimization in smart buildings. He received the Ph.D. degree in Electrical Engineering from the University of New South Wales, Sydney, Australia in 2011 and was a visiting Ph.D. scholar in the Department of Computer Science, North Carolina State University, USA in 2008. His research interests span the areas of IoT applications, cybersecurity and software defined networking. Arun has received several awards from IBM for outstanding technical accomplishments. He is the recipient of the Best Paper Award at the ACM e-Energy 2018 conference, is appointed Distinguished Speaker of ACM and is a Senior Member of IEEE.

Dan Wang (S'05, M'07, SM'13) received his B.Sc degree from Peking University, Beijing, China, in 2000, his M.Sc degree from Case Western Reserve University, Cleveland, Ohio, in 2004, and his Ph.D. degree from Simon Fraser University, Burnaby, British Columbia, Canada, in 2007, all in computer science. He is currently an associate professor at the Department of Computing, Hong Kong Polytechnic University. His research interests include network architecture and QoS, smart buildings and Industry 4.0.

Fangming Liu received the B.Eng. degree from the Tsinghua University, Beijing, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong. He is currently a Full Professor with the Huazhong University of Science and Technology, Wuhan, China. His research interests include cloud computing and edge computing, datacenter and green computing, SDN/NFV and 5G/6G. He received the National Natural Science Fund (NSFC) for Excellent Young Scholars, and the National Program Special Support for Top-Notch Young Professionals.