Reducio: Model Reduction for Data Center Predictive Digital Twins via Physics-Guided Machine Learning

Zhiwei Cao
zhiwei003@ntu.edu.sg
Nanyang Technological University, Singapore

Ruihang Wang
ruihang001@ntu.edu.sg
Nanyang Technological University, Singapore

Xin Zhou
zhouxin@ntu.edu.sg
Nanyang Technological University, Singapore

Yonggang Wen
ygwen@ntu.edu.sg
Nanyang Technological University, Singapore

ABSTRACT

The digital twin, as a digital counterpart of a physical entity, has shown great potential in data center prototyping and predictive thermal management. In this regard, Computational Fluid Dynamics/Heat Transfer (CFD/HT) models have been widely adopted. However, the computing time of the CFD/HT simulation is prohibitively long in practice. The Proper Orthogonal Decomposition (POD) has been explored to approximate the CFD/HT simulation by a linear combination of POD modes and coefficients. Existing approaches to calculating the POD coefficients use either the black-box interpolation or the simplified physical model, leading to unsatisfactory generalization ability. To advance existing approaches, this paper proposes *Reducio*, a physics-guided model reduction approach based on the POD to predict the temperature field by following two key phases of i) offline POD modes calculation and coefficients interpolation and ii) online coefficients extrapolation supervised by the principle of energy balance. To extrapolate the coefficients with limited training data, we adopt the Gaussian Process (GP) model to learn a nonlinear map between the boundary conditions and POD coefficients. We conduct two case studies in two data centers with different scales. Evaluation results in the edge data center show that Reducio achieves sub-1°C mean absolute error (MAE) in temperature field prediction compared with the CFD/HT simulation result, outperforming the existing method based on the simplified physical model by 1.5 °C. When evaluating in the industry-grade hyper-scale data center with the sensor measurements, around 1°C temperature prediction MAE is observed. Furthermore, Reducio can predict the full-fledged temperature field in *real-time*, making it a strong candidate for building data center predictive digital twins.

CCS CONCEPTS

• **Computing methodologies** → **Modeling and simulation**; • **Applied computing** → **Engineering**.

KEYWORDS

Data center, digital twin, Gaussian process, proper orthogonal decomposition, computational fluid dynamics

BuildSys '22, November 9–10, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9890-9/22/11…$15.00
https://doi.org/10.1145/3563357.3564050

ACM Reference Format:
Zhiwei Cao, Ruihang Wang, Xin Zhou, and Yonggang Wen. 2022. Reducio: Model Reduction for Data Center Predictive Digital Twins via Physics-Guided Machine Learning. In *The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '22), November 9–10, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3563357.3564050

1 INTRODUCTION

The scale of data centers has been growing continuously in recent years to support the ever-increasing cloud service demand. Meanwhile, the worldwide growth of data centers also poses substantial challenges in optimizing data center sustainability since data centers are energy-intensive enterprises. A latest survey showed that data centers contribute 0.3% global carbon emissions and the escalating trend will continue in the next decade [2]. Therefore, effective data center management is highly desirable to ensure business continuity and boost sustainability. Current data centers are usually equipped with the Data Center Infrastructure Management (DCIM) system to monitor the system states and provide information for the operators to identify potential risks such as unplanned server shutdowns, local hot spots and etc. However, with the rapid growth in data center scale and complexity, such reactive monitoring makes it difficult for the operators to forecast potential failures. In this regard, it is necessary to extend the DCIM with *accurate* and *timely* predictive models.

We consider *predictive digital twins* as a solution for proactive data center thermal management. A digital twin is the digital counterpart of a physical entity, enabling multi-physics and multi-scale simulation as well as the probabilistic modeling of the physics entity [18]. The predictive digital twin of a data center is expected to characterize the thermal and airflow distributions given certain boundary conditions. To fulfill the expectations, Computational Fluid Dynamic/Heat Transfer (CFD/HT) simulation is widely adopted [14, 19]. It can derive the fine-grained temperature field by solving the energy balance and Navier-Stokes equations. To pursue high-fidelity simulation, effective CFD/HT model calibration is important. For example, as reported in [23], the calibrated CFD/HT models can achieve sub-1°C temperature prediction errors. Although the calibrated CFD/HT model can achieve accurate prediction, the high computation overhead still limits its adoption for timely temperature prediction. For instance, when the CFD/HT simulation is conducted for hyper-scale data centers with fine mesh granularity, the solving time may vary from hours to days. Therefore, the CFD/HT simulation is usually implemented in prototyping data center design.

To facilitate timely temperature prediction, alternative surrogate models with low computation overhead will be preferred. In

Table 1: Summary of the existing works related to data center thermal modeling.

Data Center Thermal Models			Accuracy	Speed	Temperature Field	Size (m^2)
Physics-guided	High Fidelity Modeling	Turbulent CFD/HT Model [23]	MAE: sub-1 °C	hours	Yes	>800
	Simplified Modeling	Potential Flow [10]	MAE: 2.4 °C	~23 s	Yes	~500
		Heat Recirculation [7]	MAE: ~2 °C	real time	No	N/A
		Fast Fluid Dynamics [8]	NRMSE: 4%	~250 s	Yes	~660
Data-driven	Proper Orthogonal Decomposition	Flux Matching [16]	MAE: 1.24 °C	real time	Yes	~20
		Galerkin Projection [15]	MAE: 1.36 °C	~4 s	Yes	~20
		Spline Interpolation [17]	MAE: sub-1 °C	real time	Yes	~100
		Reducio	MAE: sub-1 ° C	real time	Yes	>800
	Machine Learning Approaches	Time Series Prediction [11]	MAE: ~3 °C	real time	No	30
		Gaussian Process [1]	MAE: sub-1 °C	real time	No	~50

this regard, the data-driven approaches that leverage the power of machine learning models become prevailing in recent years. For example, some researchers adopted machine learning models such as the Gaussian Process (GP) model and the Artificial Neural Network (ANN) model to learn a regression function between the boundary conditions and the temperature of a number of discrete spatial points [1]. Such an approach is straightforward and achieves satisfactory prediction accuracy. However, its flexibility and scalability are poor because one should train a prediction model for each spatial point of interest. To achieve full-fledged temperature field prediction, the Proper Orthogonal Decomposition (POD) is a good candidate technique. The idea of the POD is to express a temperature field with the linear combination of the orthogonal basis functions (i.e., POD modes) and the corresponding coefficients. The POD coefficients are determined by specific boundary conditions. To model the relationship, some researchers attempted to project the POD modes into the governing equations via Galerkin Projection and solve the high-dimensional algebraic system directly [15]. However, such a method is computationally expensive with high dimensional temperature field. To accelerate the POD coefficients calculation, the simplified heat flux matching is proposed based on the energy balance principal [16]. However, the oversimplified models may generate unsatisfactory performance, especially for areas not covered by the heat flux matching process. Other researchers proposed to build the interpolation functions between the boundary conditions and the POD coefficients, and satisfactory prediction accuracy is achieved. [17]. However, the black-box methods do not consider the underlying physical process. In other word, the predicted temperature field might violate certain physical constraints such as room-level energy balance as can be seen in §6. Therefore, an approach that incorporates physical knowledge into the POD coefficient interpolation process is still missing.

To bridge this gap, we propose *Reducio*, a physics-guided machine learning approach for building accurate and timely data center predictive digital twins. Our approach is developed based on the POD, a well-established modal analysis method in fluid dynamics [21] and the GP model [24], a powerful machine learning method that can work with limited data and quantify prediction uncertainty. Our approach consists of two stages: a) the offline POD mode computation and GP predictor construction; b) the online physics-guided temperature field prediction. In the offline phase, the POD modes are first extracted from empirical CFD/HT simulation results. For each CFD/HT simulation result, we project it into every POD mode to obtain the corresponding POD coefficient. Subsequently, the GP models are trained to map the boundary conditions to the derived coefficients. The online stage is motivated by the prediction-correction framework proposed recently in the safe reinforcement learning community [13], where a differentiable optimization layer is added to regulate the prediction result from a neural network to satisfy certain constraints. In this paper, we extend the idea to the POD coefficient calculation and formulate a constrained optimization problem to rectify the POD coefficients from the GP models. For a test case with new boundary conditions, we first leverage the trained GP models to produce *coarse* estimation of the POD coefficients. Subsequently, we *rectify* the coarse estimation in their vicinity with the energy balance principal. Finally, the temperature field is reconstructed with the linear combination of the *rectified* POD coefficients and corresponding POD modes. We evaluate the proposed method in an edge data center against CFD/HT simulation results and in an industry-grade hyper-scale data center against real temperature sensor measurement. In the former case, the proposed method achieves sub-1°C mean absolute error (MAE) and that in the latter case is around 1°C. Moreover, Reducio achieves *real-time* prediction, making it suitable for proactive data center thermal management.

We summarize our main contributions as follows.

- We propose Reducio, a physics-guided machine learning approach for CFD/HT model reduction based on the POD technique and the GP models to achieve *fast* and *accurate* full-fledged temperature field prediction.
- In Reducio, a novel post-hoc rectification method is proposed to inject energy balance principals into the POD coefficients calculation to significantly boost temperature prediction accuracy. In this regard, a convex optimization problem is formulated and optimal solution can be found numerically.
- We evaluate Reducio in two data centers with different scales extensively to show its superior prediction accuracy. When evaluating in the edge data center, Reducio can achieve around 0.5 °C MAE in the temperature field prediction, outperforming the existing POD coefficients calculation method by 1.5 °C. As for the hyper-scale data center case, about 1 °C MAE is available against the sensor measurements.

Table 2: Summary of notations.

Sym.	Definition	Sym.	Definition
$\|\|\cdot\|\|_2$	ℓ_2-norm	$\mathbf{T}^{\text{sup}}$	vector of CRAC supply temperature
$\|\|\cdot\|\|_F$	Frobenius norm	$\mathbf{V}^{\text{sup}}$	vector of CRAC volumetric air flow rates
$\langle\cdot,\cdot\rangle$	inner product	$\mathbf{P}$	vector of server power
m	number of CRACs	$\mathbf{V}$	vector of server inlet volumetric air flow rates
n	number of servers	$\mathbf{b}$	vector of boundary conditions
N	number of training cases	$\boldsymbol{\alpha}$	vector of per Watt server inlet volumetric air flow rate
r	number of truncated POD modes	$\widetilde{\mathbf{b}}$	vector of aggregated boundary conditions
D	dimension of full-fledged temperature field	$\widehat{\mathbf{a}}$	vector of coarse estimation of POD coefficients
$\mathbf{T}^{\text{obs}}$	temperature field observation dataset	$\boldsymbol{\sigma}$	vector of standard deviation of POD coefficient prediction
$\mathbf{a}$	vector of POD coefficients	$\mathbf{a}_k^{\text{tar}}$	vector of regression target for the k-th GP model
$\boldsymbol{\phi}$	vector of POD mode	C_p, ρ	specific heat and air density

2 RELATED WORK

In this section, we review the relevant studies in data center thermal modeling which are categorized into physics-driven and data-driven methods, respectively. Table 1 summarizes the existing works on data center thermal modeling.

2.1 Physics-Guided Thermal Modeling

Physics-guided thermal models are developed based on the first principal laws that govern the thermodynamics. In the past decades, the CFD/HT simulation is the representative tool to simulate the thermodynamics within a data center [14, 19, 23]. It is advantageous in deriving the full-fledged temperature field that allows the operators to perform the what-if analysis. However, the simulation of a CFD/HT model with fine-grained mesh cells can take several hours, which is prohibitive for timely prediction.

Recently, some researchers have attempted to either develop advanced numerical solvers or solve simplified physical models to accelerate the CFD/HT simulation. For example, some researchers utilized the potential flow method, which solves the simplified system with less computational efforts [10]. Even though this kind of method can achieve considerable acceleration compared with the conventional CFD/HT simulation, the simplified approximation may compromise the accuracy. Instead of solving a simplified system, Zuo *et al.* proposed the Fast Fluid Dynamics (FFD) for the sake of fast and accurate simulation of data center thermodynamics [8, 25]. The FFD method solves identical governing systems with an advanced numerical algorithm, achieving 50x acceleration. Although substantial acceleration is achieved with the FFD method, the simulation time is still at the scale of hundreds of seconds (around 250 seconds to simulate a 660 m^2 data hall [8]). To further accelerate the FFD simulation, an in situ adaptive tabulation (ISAT) algorithm is combined with the FFD simulation model [22]. For a test case, the ISAT algorithm will retrieve the simulation result from the dataset containing the offline FFD simulation results if the estimated prediction error is within a pre-defined tolerance threshold and the FFD simulation will be conducted if no matching result is found. Although further acceleration is achieved via the ISAT algorithm, it still cannot satisfy the real-time prediction requirement for an arbitrary testing case because the FFD simulation should be conducted inevitably if no matching result can be found in the oracle.

2.2 Data-Driven Thermal Modeling

To achieve real-time simulation, a number of researchers have developed data-driven approaches for data center temperature prediction. The black-box data-driven thermal modeling is to learn a function that maps the boundary conditions to the temperature at certain discrete points. Recently, Athavale *et al.* proposed to use the machine learning tools for direct temperature prediction [1]. They found that sub-1 °C temperature prediction error can be achieved with the GP regression model. However, they also found that the prediction accuracy degraded drastically with fewer training data, and thus its generalization is worrisome in practice. In contrast to the pure data-driven methods, some researchers have attempted to incorporate simplified physics models into the data-driven thermal models to improve generalization and interpretability. For instance, Li *et al.* proposed the Thermocast, a transient server temperature predictor that incorporates sensor measurements and simplified physics models, to predict the server temperature ahead of time [11]. Different from the Thermocast which focuses on transient thermal dynamic prediction, Gupta *et al.* proposed to leverage the Heat Recirculation Matrix (HRM) to establish the linear function between the steady-state server inlet temperature and server heat loads [7]. However, both approaches cannot yield satisfactory prediction accuracy.

Another thread in data-driven data center thermal modeling is based on the POD and a full-fledged temperature field can be predicted with POD-based approaches [5, 15–17]. The basic idea of the POD is to decompose a temperature field into a set of POD modes and it can be represented by their linear combination. The POD modes can first be extracted from empirical CFD/HT simulation results by solving an eigenvalue problem and the function between the coefficients in the linear combination and boundary conditions are established subsequently to deal with new test cases. As for the mapping function, some researchers establish an algebraic system to derive the POD coefficients for a test case. The algebraic system can be established either from the Galerkin Projection [5, 15] or the heat flux matching process [16]. A high dimensional algebraic system is built from Galerkin Projection and its solving time is significantly longer than the simplified heat flux matching counterpart. For the heat flux matching-based approach, a linear system relating the POD coefficients and the boundary conditions is established and the least square technique is utilized to derive the POD

coefficients. Other researchers utilized the interpolation methods to directly learn a function that maps boundary conditions into POD coefficients without physics knowledge [17]. These methods are reported to have satisfactory prediction accuracy. However, the interpolation does not involve physics knowledge, rendering it unreliable when dealing with varying boundary conditions and complicated 3D geometry.

3 PRELIMINARY

In this section, we provide necessary preliminary knowledge. We first introduce the POD method. The GP regression model is briefly introduced in the following. The notations used in this paper are listed in Table 2.

3.1 Proper Orthogonal Decomposition

The POD, also known as the Karhunen-Loeve decomposition [21], is a powerful tool for designing surrogate models for complex multi-scale turbulent convective systems [15]. The basic idea of POD is to approximate the temperature field by the linear combination of POD modes that capture the coherent structure in the temperature field. It has been widely used for modal analysis in the fluid dynamics community [21]. In this paper, we consider a finite dimension steady-state temperature field as $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^D$ where $\mathbf{x}$ is the spatial coordinate in the cartesian coordinate system and D is the number of grid cells in CFD/HT simulation. $\mathbf{T}(\mathbf{x})$ can be expanded with the POD modes:

$$\mathbf{T}(\mathbf{x}) = \sum_i a_i \boldsymbol{\phi}_i(\mathbf{x}), \tag{1}$$

where a_i is the POD coefficient for the i-th POD mode $\boldsymbol{\phi}_i(\mathbf{x}) \in \mathbb{R}^D$. To efficiently solve the POD modes, we leverage the Method of Snapshots proposed by Sirovich *et al.* [20]. Specifically, we can obtain N temperature field $\mathbf{T}_i^{\text{obs}}(\mathbf{x}),\ i = 1, 2, ..., N$ by running N CFD/HT simulation with *different* boundary conditions, which forms a training dataset $\mathbf{T}^{\text{obs}} = [\mathbf{T}_1^{\text{obs}}(\mathbf{x}), \mathbf{T}_2^{\text{obs}}(\mathbf{x}), ..., \mathbf{T}_N^{\text{obs}}(\mathbf{x})] \in \mathbb{R}^{D\times N}$ where each column is a simulated temperature field. Formally, the POD generates a set of basis that forms a low-rank matrix $\boldsymbol{\Phi} \in \mathbb{R}^{D\times r}$ so that the following objective function is minimized:

$$\underset{\boldsymbol{\Phi},\ \text{s.t. rank}(\boldsymbol{\Phi})=r}{\text{argmin}} \quad ||\mathbf{T}^{\text{obs}} - \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal \mathbf{T}^{\text{obs}}||_F. \tag{2}$$

where the i-th column of $\boldsymbol{\Phi}$ is the i-th POD mode $\boldsymbol{\phi}_i$ and r is the rank of the truncation. This optimization problem can be solved by implementing the Singular Value Decomposition (SVD) on the training dataset $\mathbf{T}^{\text{obs}}$. After we obtain the POD modes, the key challenge is to determine the POD coefficients for a new case, which will be addressed in §5.

3.2 Gaussian Process Regression

In this section, we briefly introduce the GP model and its application in the regression problem. We refer interested readers to [24] for more details.

Consider a noisy observation y from an underlying function $f : \mathbb{R}^n \to \mathbb{R}$ through a Gaussian noise model $y = f(\mathbf{x}) + n$ where $\mathbf{x} \in \mathbb{R}^n$ is the feature vector and n is the Gaussian noise with distribution $\mathcal{N}(0, \sigma_n^2)$. The statistical property of y is fully specified by its mean function $\mu(\mathbf{x}; \theta_\mu) = \mathbb{E}\left[f(\mathbf{x}); \theta_\mu\right]$ and covariance function

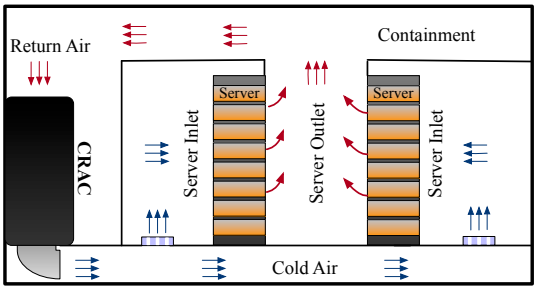

Figure 1: Illustration of an air-cooled raised-floor data center with hot-aisle containment.

$\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j; \theta_\mathcal{K}) = \mathbb{E}\{[f(\mathbf{x}_i) - \mu(\mathbf{x}_i; \theta_\mu)] \cdot [f(\mathbf{x}_j) - \mu(\mathbf{x}_j; \theta_\mu)]; \theta_\mathcal{K}\}$ where θ_μ and $\theta_\mathcal{K}$ are the hyperparameters of the GP model.

The application of GP in regression problem is straightforward. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^T \in \mathbb{R}^{N\times p}$ be the feature vector set containing N samples where $\mathbf{x}_i \in \mathbb{R}^p$ is the i-th feature vector and $\mathbf{Y} = [y_1, y_2, ..., y_N]^T \in \mathbb{R}^N$ be the label set. Let $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ be the training dataset. It is important to note that the GP is a *nonparametric* model which means that in the training phase, we only need to store the training dataset $\mathcal{D}$ and estimate the hyperparameters θ_μ and $\theta_\mathcal{K}$ via Maximum Likelihood Estimation (MLE) [24]. In the inference phase, let us consider a new test sample $\mathbf{x}^* \in \mathbb{R}^n$ and denote the output of the GP model as y^* which follows a Gaussian distribution $\mathcal{N}(\bar{y}_*, \sigma_*^2)$ where the $\bar{y}_*$ and the σ_*^2 can be computed in the following way:

$$\bar{y}_* = \mu(\mathbf{x}_*; \theta_\mu^*) + \mathcal{K}_*\mathcal{K}^{-1}\left[\mathbf{Y} - \mu(\mathbf{X}; \theta_\mu^*)\right], \tag{3}$$

$$\sigma_*^2 = \mathcal{K}_{**} - \mathcal{K}_*\mathcal{K}^{-1}\mathcal{K}_*^T, \tag{4}$$

Here, $\mathcal{K}_* = [\mathcal{K}(\mathbf{x}^*, \mathbf{x}_1; \theta_\mathcal{K}^*), ..., \mathcal{K}(\mathbf{x}^*, \mathbf{x}_N; \theta_\mathcal{K}^*)] \in \mathbb{R}^N$ and $\mathcal{K} \in \mathbb{R}^{N\times N}$ is the covariance matrix.

4 SYSTEM MODEL

In this section, we first present the system configuration. Subsequently, we introduce the multi-scale data center thermal modeling.

4.1 System Configuration

In this paper, we consider a data center with m Computer Room Air Conditioning (CRAC) units and n servers. The layout of a typical air-cooled data center with hot aisle containment is illustrated in Fig. 1. The i-th CRAC unit is specified by its air supply temperature T_i^{sup} (°C) and supply air volumetric flow rate V_i^{sup} (m^3/s). The configuration of a server includes its power consumption and inlet air volumetric flow rate.

Inputs: The input to the predictive digital twin are a vector containing all boundary conditions and r POD modes. Specifically, the boundary condition vector is $\mathbf{b} = (\mathbf{T}^{\text{sup}}, \mathbf{V}^{\text{sup}}, \mathbf{P}, \mathbf{V})$ where $\mathbf{T}^{\text{sup}} = [T_1^{\text{sup}}, T_2^{\text{sup}}, ..., T_m^{\text{sup}}]$, $\mathbf{V}^{\text{sup}} = [V_1^{\text{sup}}, V_2^{\text{sup}}, ..., V_m^{\text{sup}}]$, $\mathbf{P} = [P_1, P_2, ..., P_n]$, $\mathbf{V} = [V_1, V_2, ..., V_n]$ are the vector of air supply temperatures, the vector of supply air volumetric flow rates, the vector of server heat loads, and the vector of server inlet air volumetric flow rates, respectively.

Outputs: The result is a *full-fledged* temperature field $\widetilde{\mathbf{T}}(\mathbf{x}) = \sum_{i=1}^{r} g_i(\mathbf{b})\boldsymbol{\phi}_i(\mathbf{x})$, where the function $g_i(\mathbf{b})$ maps the boundary condition vector $\mathbf{b}$ into the coefficient for the i-th POD mode.

4.2 Multi-Scale Data Center Thermal Modeling

Data centers are representative *multi-scale* turbulent convective system [15]. In other words, we need to consider the energy balance from the local individual server to the global room space. In the server level, let the inlet and outlet temperature of the i-th server be T_i^{in} and T_i^{out}, the server level local energy balance is written as:

$$P_i = C_p \rho V_i (T_i^{\text{out}} - T_i^{\text{in}}), \tag{5}$$

where V_i is the i-th server inlet air volumetric flow rate, ρ is the air density and C_p is the specific heat. In this paper, we model V_i as a linear function of P_i by:

$$V_i = \alpha_i \cdot P_i, \tag{6}$$

where α_i is the per Watt server inlet air volumetric flow rate ($\text{m}^3/(\text{s}\cdot\text{W})$) for the i-th server [12]. In general, α_i cannot be obtained from the server hardware specification and can only be estimated via model calibration or on-site measurement [23]. In this paper, we assume that all α_i are known as a prior.

In the room level, we adopt a nodal model and view the entire data hall as a node similar to the treatment in EnergyPlus [3]. The room level energy balance is expressed as:

$$\sum_{i=1}^{m} C_p \rho V_i^{\text{sup}} (T_i^{\text{ret}} - T_i^{\text{sup}}) = \sum_{k=1}^{n} P_i, \tag{7}$$

where T_i^{ret} is the return air temperature of the i-th CRAC unit.

5 THE REDUCIO APPROACH

In this section, we first provide an overview of the proposed Reducio approach. Subsequently, we elaborate on our solution in detail.

5.1 Approach Overview

The workflow of Reducio is illustrated in Fig. 2 and it can be divided into the offline phase and online phase. In the offline phase, we first conduct multiple CFD/HT simulations under different boundary conditions to obtain the observation dataset $\mathbf{T}^{\text{obs}}$. This step is the most time-consuming one in the entire workflow since multiple CFD/HT simulations are implemented. After we obtain the observation dataset, the Method of Snapshot introduced in §3.1 is applied to obtain the POD modes. Following that, we project each temperature field into the POD modes to derive their corresponding POD coefficients, which will serve as the labels for the GP models. Subsequently, we construct GP models that learn the mapping between the boundary condition and the derived POD coefficient. Since the common GP model cannot make vector prediction, we ignore the correlation between POD coefficients and train one GP model for *each* POD coefficient [1].

In the online phase, the core challenge is to infer the POD coefficient of each POD mode given *new* boundary conditions. In this regard, we design a novel physics-guided *two-step* framework. To be specific, we first leverage the offline trained GP models to provide a coarse estimation of each POD coefficient given the new boundary conditions. After that, based on the POD coefficients from the GP models, a physics-guided rectification is conducted within the *vicinity* of the coarse estimations, which will be addressed in detail in §5.3.2. Finally, the temperature field is predicted by the linear combination of the POD modes and the rectified POD coefficients.

[1]Vector prediction with the GP model is also feasible, but we empirically found that treating POD coefficients as independent random variables is sufficient.

5.2 Offline GP Model Training

In this paper, the GP regression model is leveraged to learn a nonlinear mapping between boundary conditions and each POD coefficient. The reasons for choosing GP models are three folds. Firstly, according to the POD theory, each case in the training dataset can be reconstructed with the extracted POD modes and the coefficient for each POD mode is the inner product of the training case and the POD modes. By using the GP model, we can guarantee that if the testing input is identical to a training input, the prediction will become the simple retrieval of the corresponding derived POD coefficients and the prediction will be accurate because the GP regression can be viewed as interpolation. Otherwise, the predicted POD coefficients will be the weighted average of all POD coefficients in the training dataset and the training inputs which are close to the testing input will have more impacts on the prediction result. Such a similarity-guided weighted averaging can be viewed as a *soft* look up table query and it can be connected to the surrogate model based on the ISAT algorithm, where the *hard* look up table query is implemented [22]. Secondly, GP models offer good generalization ability in the *small data regime*. The CFD/HT simulation is very time-consuming, and thus it is prohibitive to obtain a large volume of simulation results to train a sophisticated machine learning model such as the deep neural network. Therefore, GP models are preferable when dealing with simulation data. Thirdly, GP models provide the *uncertainty* for the prediction. The uncertainty information is helpful to define the feasible search region for the proposed physics-guided rectification, which will be discussed in §5.3.2.

For r POD modes used to predict the temperature field, we establish r GP regression models, i.e., one for each POD mode, and the input as well as the regression target for each model are specified in the following. We adopt the *aggregated* boundary conditions $\widetilde{\mathbf{b}} = [T_1^{\text{sup}}, ..., T_m^{\text{sup}}, V_1^{\text{sup}}, ..., V_m^{\text{sup}}, \sum_i P_i, \sum_i V_i] \in \mathbb{R}^{2m+2}$ as the input to each GP model. The reasons for using the aggregated boundary condition instead of the raw boundary conditions are two-fold. Firstly, a data center may host hundreds even thousands of servers, leading to the curse of dimensionality. Such a high dimensional input space will impede the learning of the GP model given limited CFD/HT simulation results, e.g., tens of samples in our cases. Secondly, the GP models are only used to generate a *coarse* estimation of the POD coefficients, and therefore the aggregated boundary condition is sufficient for this purpose. We project all temperature fields in the training dataset to the corresponding POD mode to obtain the regression target. Specifically, for the i-th GP model, to obtain the regression target for the k-th case in the training dataset, we compute the inner product of $\mathbf{T}_k^{\text{obs}}$ and the $\boldsymbol{\phi}_i$ as the target:

$$a_{k,i}^{\text{tar}} = < \boldsymbol{\phi}_i, \mathbf{T}_k^{\text{obs}} >, \quad k = 1, 2, ..., N, \tag{8}$$

As for the mean and kernel functions of the GP model, we use the widely adopted constant mean function $\mu(\tilde{\mathbf{b}}) = 0$ and the squared

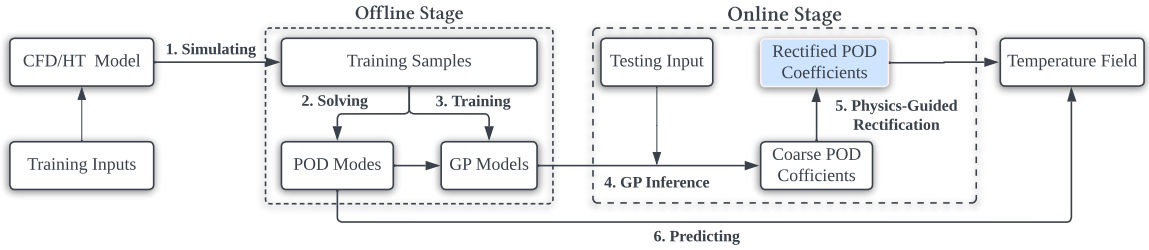

Figure 2: Workflow of Reducio. In the offline phase, POD modes are extracted from the observation dataset. GP models which learn the mapping between the boundary conditions and the POD coefficients are trained subsequently. In the online phase, we obtain the coarse estimation of each POD coefficient with the corresponding GP model and rectify the estimation by the physics-guided rectification. Finally, the rectified POD coefficients are used for temperature field prediction.

exponential kernel $\mathcal{K}(\widetilde{\mathbf{b}}_i, \widetilde{\mathbf{b}}_j) = \sigma^2 \exp\left(-\frac{||\widetilde{\mathbf{b}}_i - \widetilde{\mathbf{b}}_j||_2}{2l^2}\right)$ where σ and l are two hyperparameters and they can be optimized by GPy [6]. It should also be noted that all GP models share the same form of the mean and kernel function but with different hyperparameters.

5.3 Online POD Coefficient Estimation

In this section, we describe the online stage of Reducio, a physics-guided two-step approach for estimating the POD coefficients given *new* boundary conditions. We first introduce how to obtain a coarse estimation of POD coefficients with the offline optimized GP models. Following that, we will provide the formulation of the physics-guided rectification.

5.3.1 Coarse Estimation with GP Models. Given a test case with aggregated boundary condition vector $\widetilde{\mathbf{b}}_*$ and r optimized GP models, we produce r coarse estimation of the coefficients for the corresponding POD modes. Specifically, for the i-th POD coefficient, we use the posterior mean defined in Eq. (3) as the coarse estimation. By repeating this process for r times, we can obtain the coarse estimation of r POD coefficients, denoted as $\widehat{\mathbf{a}} = [\hat{a}_1, ..., \hat{a}_r]$. More importantly, we can also obtain the prediction uncertainty from Eq. (4), and the result forms a vector defined as $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, ..., \sigma_r]$. However, it should be noted that such a procedure does not guarantee that the predicted temperature field will respect the global and local energy balance without imposing explicit constraints. This motivates us to design the physics-guided local search based on the coarse estimation $\widehat{\mathbf{a}}$ and the uncertainty vector $\boldsymbol{\sigma}$ produced by the GP models so that local and global energy balance can be incorporated in the POD coefficient calculation.

5.3.2 Rectification with Physics-Guided Local Search. In this section, we formulate a constrained optimization problem so that the multi-scale energy balance is incorporated in the inference process. We first plug the POD expansion of a temperature field into the Eq. (5) as:

$$P_i = C_p \rho V_i \Big\{ \sum_{k=1}^{r} a_k \left[\phi_k(\mathbf{x}_i^{\text{out}}) - \phi_k(\mathbf{x}_i^{\text{in}}) \right] \Big\}. \tag{9}$$

By finding a proper set of POD coefficients in Eq.(9), the local energy balance of the i-th server will be satisfied in the predicted temperature field. Similarly, we can also incorporate the POD expansion into the room level energy balance Eq. (7) as:

$$\sum_{i=1}^{m} C_p \rho V_i^{\text{sup}} \Big\{ \sum_{k=1}^{r} a_k \phi_k(\mathbf{x}_i^{\text{ret}}) - T_i^{\text{sup}} \Big\} = \sum_{k=1}^{n} P_i, \tag{10}$$

Based on Eq. (9) and Eq. (10), the rectification aims to search for a POD coefficient vector $\mathbf{a}$ in the *vicinity* of the coarse estimation $\widehat{\mathbf{a}}$ so that the global energy balance will be satisfied while the local energy balance is satisfied as much as possible. The reason why we cannot make all local energy balance hold is that the number of servers hosted in a data center will exceed the number of POD modes by a large margin, i.e., $n >> r$ in practice. Therefore it is unlikely to find a set of POD coefficients that satisfy all local energy balance simultaneously. Meanwhile, it should be noted that there exists only one global energy balance to be satisfied and thus we can enforce its satisfaction. Guided by these modeling principles, we can formulate the physics-guided local search problem as a Quadratic Constrained Quadratic Programming (QCQP) in the following:

$$\begin{aligned} \min_{\mathbf{a}} \quad & \|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \\ \text{s.t.} \quad & \sum_{i=1}^{n} \Big\{ \frac{P_i}{C_p \rho V_i} - \mathbf{a}^\intercal \left[\boldsymbol{\Phi}(\mathbf{x}_i^{\text{out}}) - \boldsymbol{\Phi}(\mathbf{x}_i^{\text{in}}) \right] \Big\}^2 \le \epsilon, \\ & \sum_{i=1}^{m} C_p \rho V_i^{\text{sup}} \left[\mathbf{a}^\intercal \boldsymbol{\Phi}(\mathbf{x}_i^{\text{ret}}) - T_i^{\text{sup}} \right] = \sum_{k=1}^{n} P_i, \\ & \hat{a}_i - \beta\sigma_i \le a_i \le \hat{a}_i + \beta\sigma_i, \quad i = 1, 2, ..., r. \end{aligned} \tag{11}$$

Here, $\boldsymbol{\Phi}(\mathbf{x}_i^{\text{out}}) = [\phi_1(\mathbf{x}_i^{\text{out}}), ..., \phi_r(\mathbf{x}_i^{\text{out}})] \in \mathbb{R}^r$ contains the values of r POD modes at the i-th server outlet. $\boldsymbol{\Phi}(\mathbf{x}_i^{\text{in}})$ and $\boldsymbol{\Phi}(\mathbf{x}_i^{\text{ret}})$ are defined similarly. It should also be noted that the local search is also bounded by the uncertainty of each POD coefficient, which is produced by the GP models. If the uncertainty is large, we should count more on the multi-scale energy balance to obtain a reasonable estimation of the POD coefficients and vice versa. Therefore, the GP model offers unique advantages over other machine learning methods in the physics-guided rectification. As for the determining of ϵ in the first constraint of optimization problem (11), we plug in the coarse estimation result $\widehat{\mathbf{a}}$ into the LHS of the first constraint, i.e., $\epsilon = \sum_{i=1}^{n} \Big\{ \frac{P_i}{C_p \rho V_i} - \widehat{\mathbf{a}}^\intercal \left[\boldsymbol{\Phi}(\mathbf{x}_i^{\text{out}}) - \boldsymbol{\Phi}(\mathbf{x}_i^{\text{in}}) \right] \Big\}^2$. By imposing such a constraint, we aim to make the rectified POD coefficients produce

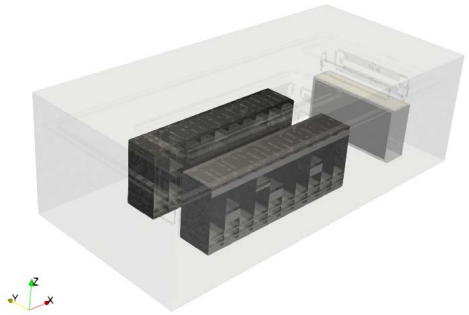

Figure 3: Layout of the studied edge data center. There exists one CRAC unit and two racks which host 70 homogeneous servers in the data hall. Sensors are placed at server inlets and outlets, as well as CRAC unit return.

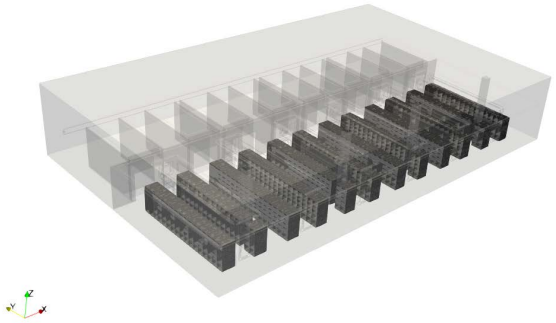

Figure 4: Layout of the studied hyper-scale data center. It hosts near 2000 servers and 26 temperature sensors are installed at the hot and cold aisles to record the temperature.

smaller local energy balance violation than that incurred by the coarse estimation from the GP models.

We solve the QCQP by problem reformulation. Specifically, we introduce an auxiliary variable t so that the problem can be reformulated as:

$$\begin{aligned} \min_{\mathbf{a},t} \quad & t \\ \text{s.t.} \quad & \sum_{i=1}^{n} \left\{ \frac{P_i}{C_p \rho V_i} - \mathbf{a}^\intercal \left[\mathbf{\Phi}(\mathbf{x}_i^{\text{out}}) - \mathbf{\Phi}(\mathbf{x}_i^{\text{in}}) \right] \right\}^2 \leq \epsilon \\ & \sum_{i=1}^{m} C_p \rho V_i^{\text{sup}} \left[\mathbf{a}^\intercal \mathbf{\Phi}(\mathbf{x}_i^{\text{ret}}) - T_i^{\text{sup}} \right] = \sum_{k=1}^{n} P_i, \\ & ||\mathbf{a} - \widehat{\mathbf{a}}||_2 \leq t, \\ & \hat{a}_i - \beta\sigma_i \leq a_i \leq \hat{a}_i + \beta\sigma_i, \quad i = 1, 2, ..., r. \end{aligned} \tag{12}$$

We can identify that the optimization problem Eq. (12) is a Second Order Cone Programming (SOCP) which can be solved efficiently by CVXPY [4]. As for the β in the last constraint of the optimization problem Eq. (12), we start with $\beta = 1$, and if the optimization problem is infeasible, we multiply it by a constant until a feasible solution is found[2].

6 EVALUATION

6.1 Evaluation Methodology

In this section, we first present the metric used in this paper for performance evaluation. Subsequently, we introduce the evaluated baseline approaches. Evaluation results on two data centers are discussed in the following.

6.1.1 Evaluation Metric. We evaluate different approaches from these two perspectives: a) temperature prediction error and, b) computation overhead. In terms of the temperature prediction error, we use the temperature field generated by the CFD/HT model as the ground truth unless particularly specified. We use the MAE metric to evaluate temperature prediction error, which is defined as $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\widetilde{T}(\mathbf{x}_i) - T(\mathbf{x}_i)|$ where N is the number of evaluation point, $\mathbf{x}_i$ is the spatial coordinate of the i-th evaluation point, and $\widetilde{T}(\mathbf{x}_i)$ and $T(\mathbf{x}_i)$ are the predicted temperature and the ground truth temperature, respectively. As for the computation overhead, we evaluate the computing time of different models running with the same hardware configuration and computation resource.

6.1.2 Baseline Approaches. In this section, we introduce the baseline approaches that are compared against Reducio.

- **POD-FluxMatching** (POD-FM) [16]. This method utilizes the local energy balance and global energy balance to jointly form a linear system and calculate the POD coefficients by solving the linear system with the least square technique. This approach finds the minimizer of the summation of Eq. (9) and Eq. (10).
- **GP** [1]. In contrast to the POD-based approaches, this approach cannot obtain the full-scale temperature field. For a fair comparison, we also use the aggregated boundary condition defined in §5.2 to obtain the GP predictors for the temperature of the interested spatial locations. We train one GP model for predicting the temperature of an interested spatial location.
- **POD-GP**. This method is a simplified version of Reducio which omits the POD coefficients rectification and utilizes the coarse estimation results from the offline optimized GP models directly to predict the temperature field. The prediction process can be viewed as a black-box POD coefficient interpolation.

6.2 Evaluation on Edge Data Center

The data center considered in the first case study is an edge data center and its layout is illustrated in Fig. 3. There exists one CRAC unit and two rows of racks that host 70 *homogeneous* servers. The data center is equipped with hot aisle containment to prevent hot air recirculation. We place a pair of sensors near the inlet and outlet of each server in the CFD/HT model to measure their temperatures, respectively. We also place one sensor near the CRAC unit return aisle to measure the return temperature. In this case study, we compare the estimated temperature field against the CFD/HT simulation results. Specifically, the geometry of the data hall is

[2]We empirically found that for most cases, we can obtain a feasible solution within 10 trails with the multiplication factor 2.

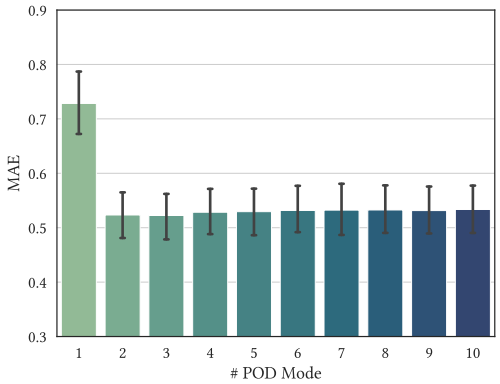

Figure 5: Convergence of the proposed method in terms of different number of POD modes. The MAE of the temperature field prediction converges with 5 POD modes.

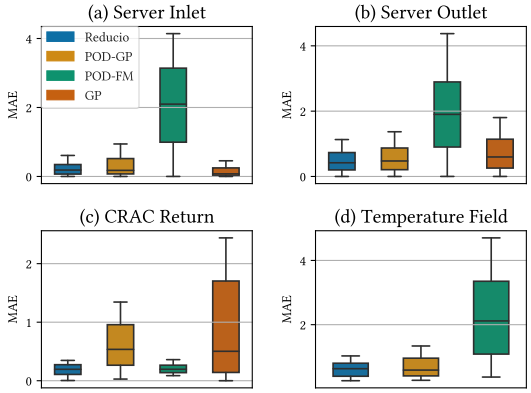

Figure 6: Comparison of temperature prediction accuracy of Reducio and other baseline methods. We evaluate the MAE for the temperature prediction result at the server inlet points, server outlet points and the whole temperature field.

meshed by OpenFOAM [9]. With the fine-grained mesh files, we call the OpenFOAM solver to derive the temperature field of the modeled space.

With the CFD model, we conduct simulations with different boundary conditions to construct the synthetic training and test dataset. Specifically, we vary the supply air temperature, supply air volumetric flow rate, and the heat load of each server. For the training dataset, the supply air temperature is chosen from $\{17, 21, 25\}$. To reduce the design space, we assume that the heat load of each server is *identical* and the heat load of each server comes from the set $\{500, 1000\}$. The per Watt server inlet air volumetric flow rate in Eq. (6) is specified as 10^{-4}. To mimic the real-world scenarios, the server inlet air volumetric flow rate is added with Gaussian noise of zero mean. The standard deviation of the Gaussian noise is the server inlet air volumetric flow rate multiplies by 0.05 to ensure an identical signal-to-noise ratio for all cases. With the total server flow rate calculated as $\bar{V} = \sum_i V_i$, we sample 5 supply air volumetric flow rates uniformly between $1.4\bar{V}$ and $1.8\bar{V}$ to guarantee that the supply air volumetric flow rate is larger than the total server inlet air volumetric flow rate. The boundary conditions for all cases are the Cartesian product of the four sets, which generates 30 synthetic training samples in total. As for the test dataset, the supply air temperature comes from the set $\{17, 18, 19, 20, 21, 22, 23, 24, 25\}$ and the server heat load is from the set $\{200, 500, 800\}$. The sampling method for the supply air volumetric flow rate is identical to that applied in generating training cases. With these settings, we generate 81 synthetic test samples in total. As for the number of PODs used in the online prediction, we show the full-scale temperature prediction MAE versus the number of used POD modes in Fig. 5. The error bar stands for the standard deviation. It is seen that MAE converges with the first five POD modes. Therefore, we leverage them for online temperature prediction.

6.2.1 Temperature Prediction Accuracy Evaluation. In this section, we compare the temperature prediction error of the proposed method and the baseline methods. For all POD-based methods, the number of POD modes used in the temperature prediction is

Table 3: Comparison of the room and server-level energy balance violation with and without rectification.

	POD-GP (w/o.)	Reducio (w/.)
Room Level	5.570 (±0.104)	**2.758e-12** (±2.158e-11)
Server Level	0.615 (±0.104)	**0.564** (±0.106)

derived from the empirical results which give the lowest prediction error. As for the GP method, we report the temperature prediction error at the local points of interest, i.e., server inlet/outlet and the CRAC unit return point. The evaluation result is presented in Fig. 6 and several conclusions can be drawn.

Firstly, we can see that the POD-GP and Reducio outperform the POD-FM significantly, showing the limitation of the simplified physical model in POD coefficient calculation. Specifically, the POD-FM approach simply builds a linear system between the boundary conditions and POD coefficients and find the least square solution for the linear system. Even though the least square solution can be derived using the simplified physical model, it might not be optimal in terms of temperature field prediction because the satisfaction of the server level and room level energy balance is the necessary but insufficient condition for temperature prediction. Instead of relying on server and room level energy balance to calculate the POD coefficients, both the POD-GP and the Reducio take advantage of the fact that similar boundary conditions will produce similar temperature field and similar POD coefficients. By weighted averaging the derived POD coefficients in the training dataset, more reliable POD coefficient prediction is expected.

Furthermore, the comparison between the POD-GP and Reducio illustrates the benefits of the rectification. As shown in Fig. 6, Reducio outperforms the POD-GP for the CRAC unit return temperature prediction. It is due to the reduced room-level energy balance violation, making the predicted temperature at the CRAC return side respect the physical truth. To validate this claim, we evaluate the room-level and server-level energy balance violation of the POD-GP

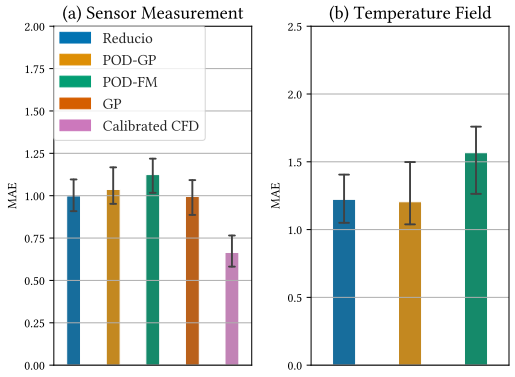

Figure 7: Comparison of temperature prediction accuracy of Reducio and other baselines for the hyper-scale data center.

and Reducio. The room-level energy balance violation is measured by $\left|\sum_{i=1}^{n}\frac{P_i}{\rho C_p}-\sum_{i=1}^{m}V_i^{\text{sup}}\left[T_i^{\text{sup}}-T_i^{\text{sup}}\right]\right|$. The server-level energy balance violation is measured by $\sqrt{\sum_{i=1}^{N}\{\frac{P_i}{\rho C_p}-[T_i^{\text{out}}-T_i^{\text{in}}]\}^2}$. The evaluation results are shown in Table. 3. We can clearly see that the room-level energy balance violation of the predicted temperature field with the POD-GP is significantly larger. By introducing the post-rectification process, the room-level energy balance violation can be close to zero because we explicitly formulate it as an equality constraint in the optimization problem of Eq. (12). We would like to highlight that the CRAC unit return temperature is the linkage between cooling energy consumption and the data hall thermodynamics if we want to implement the thermal-energy co-simulation.[3] To be specific, upon receiving the CRAC return temperature, the energy model can simulate the cooling energy consumption of the associated chiller plant. If the CRAC unit return temperature prediction is erroneous, it will compromise the accuracy of the energy model. Thus, Reducio is more suitable in the co-simulation scenario than the POD-GP. Meanwhile, the server-level energy balance violation is also slightly reduced via the rectification, which also explains the improvement for the server inlet/outlet temperature prediction by Reducio.

Thirdly, Reducio outperforms the GP baseline in terms of accuracy and flexibility. The GP approach performs significantly worse than Reducio for the CRAC unit return temperature prediction because the room energy balance is not considered explicitly in the GP approach. Furthermore, we cannot obtain the temperature field with the GP approach and 141 (70 for server inlets, 70 for server outlets and one for CRAC unit return point) GP models should be established in the case study. On the other hand, establishing 5 GP models is sufficient for Reducio to predict the temperature field with satisfactory accuracy.

6.3 Evaluation on Hyper-Scale Data Center

In this section, we evaluate the proposed method on a hyper-scale industry-grade data center which hosts thousands of servers and tens of CRACs and sensors [4], as shown in Fig. 4. We use the method in [23] to calibrate the CFD/HT model automatically to obtain the per Watt server inlet air volumetric flow rate in Eq. (6) based on historical sensor measurements. The training samples consist of 10 cases generated by running the CFD/HT with historically collected boundary conditions. We also add Gaussian noise in the server inlet air volumetric flow rate similar to that in the previous case study. The test samples consist of another 10 cases generated by running the CFD/HT simulation with *calibrated* boundary conditions.

[3]See the engineering reference of EnergyPlus for more details on how the CRAC unit return temperature can be used to link air dynamics model and the energy model for the thermal-energy co-simulation. (link: https://bit.ly/3C5i2tP, access date: 2022-9-25)

Table 4: Computing time comparison between CFD/HT simulation and data-driven surrogate models.

	Approach	Computing Time (s)
Edge Data Center	Reducio	0.08(±0.06)
	POD-GP	0.08 (±0.06)
	POD-FM	0.01 (±0.0004)
	GP	0.07 (±0.01)
	CFD/HT	283.15 (±10.12)
Hyper-scale Data Center	Reducio	0.33 (±0.01)
	POD-GP	0.23 (±0.0005)
	POD-FM	0.23 (±0.0008)
	GP	0.01 (±0.0004)
	CFD/HT	25232.12 (±32.12)

6.3.1 Temperature Prediction Accuracy Comparison. The temperature prediction accuracy evaluation result is illustrated in Fig. 7 and several conclusions can be drawn accordingly. Firstly, the POD-FM method does not achieve satisfactory temperature prediction accuracy. The reason is that the target data center has complicated 3D geometry, which will result in irregular airflow patterns and the heat flux matching model is oversimplified for dealing with such a scenario. We also find that Reducio has lower temperature prediction error compared with the POD-GP method. We believe such an improvement comes from the physics-guided rectification which injects the server and room level energy balance to the POD coefficient interpolation process. Furthermore, even though the direct GP model can also achieve similar prediction accuracy for sensor temperature measurement prediction compared with our approach, it scalability limits its practical implementation in monitoring the server temperature due to the large volume of servers hosted in the data hall. On the contrary, Reducio can still achieve around 1.0 °C MAE when compared against sensor measurements and around 1.2 °C in temperature field prediction in such a complex scenario with only ten training cases, showing its great potential in predictive data center thermal management.

6.4 Computing Time Comparison

We illustrate the computing time of different models for the two data centers in Table 4. Firstly, it is seen that all surrogate models achieve significant acceleration compared with the CFD/HT simulation, which facilitate timely thermodynamic simulation in practice. Secondly, by comparing the computing time of the POD-based method and the GP approach, we can see that the GP method is more advantageous in the hyper-scale data center case because

[4]the detailed information are omitted due to commercial interest.

we do not intend to obtain all server inlet/outlet temperatures in this case, which significantly reduces its workload. However, due to the huge mesh size of the hyper-scale data center, the POD-based method takes longer time in reconstructing the whole temperature field. It should also be pointed out that all POD-based methods can still accomplish computation within one second, and they can also be viewed as *real-time* predictive models. Lastly, Reducio has the largest computational overhead among these approaches since the physics-guided rectification involves iterative numerical optimization. Therefore, Reducio trades reasonable computational overhead with significant prediction accuracy improvement.

7 CONCLUSION AND DISCUSSION

In this paper, we propose Reducio, a physics-guided machine learning approach for CFD/HT model reduction to facilitate the development of timely and accurate predictive data center digital twins. Our approach is based on the POD technique and thus it can predict the whole temperature field instead of the temperature in discrete spatial locations. By adopting the GP model, a nonlinear mapping between boundary conditions and POD coefficients is established, offering more powerful prediction capability than previous works based on simplified physical models. Moreover, multi-scale energy balance is introduced to perform rectification on the coarse prediction from the GP models so that the predicted temperature field will respect the energy balance in a best effort manner. We conduct two case studies to evaluate the feasibility and scalability of our approach, one in an edge data center and another in a hyper-scale data center. Sub-1°C MAE is observed in the edge data center evaluation, outperforming the previous method based on simplified physical model by 1.5 °C. Moreover, around 1°C MAE is also observed in the more challenging hyper-scale data center case Furthermore, Reducio can predict the temperature field in real-time, achieving tens of thousands of acceleration compared with the CFD/HT simulation. We believe that the Reducio, can greatly facilitate learning-based control of the data center cooling system, where the intelligent agent should interact with the simulator for sufficient times to perform optimal control. The high accuracy of the Reducio can reduce the risk of online deployment of the learning-based controller when it is trained with the Reducio. Furthermore, the real-time prediction will accelerate the learning process significantly, making it possible for learning-based optimal control. In the future, we will attempt to design the safe and energy-efficient control policy to improve data center energy efficiency and sustainability with the Reducio.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Energy Research Testbed and Industry Partnership Funding Initiative of the Energy Grid (EG) 2.0 programme and its Central Gap Fund ("Central Gap" Award No. NRF2020NRF-CG001-027) and and its Sustainable Tropical Data Centre Testbed programme (STDCT).

REFERENCES

[1] Jayati Athavale, Minami Yoda, and Yogendra Joshi. 2019. Comparison of data driven modeling approaches for temperature prediction in data centers. *International Journal of Heat and Mass Transfer* 135 (2019), 1039–1052.

[2] Zhiwei Cao, Xin Zhou, Han Hu, Zhi Wang, and Yonggang Wen. 2022. Toward a Systematic Survey for Carbon Neutral Data Centers. *IEEE Communications Surveys & Tutorials* 24, 2 (2022), 895–936. https://doi.org/10.1109/COMST.2022.3161275

[3] Drury B Crawley, Linda K Lawrie, Frederick C Winkelmann, Walter F Buhl, Y Joe Huang, Curtis O Pedersen, Richard K Strand, Richard J Liesen, Daniel E Fisher, Michael J Witte, et al. 2001. EnergyPlus: creating a new-generation building energy simulation program. *Energy and buildings* 33, 4 (2001), 319–331.

[4] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[5] Rajat Ghosh and Yogendra Joshi. 2011. Dynamic reduced order thermal modeling of data center air temperatures. In *International Electronic Packaging Technical Conference and Exhibition*, Vol. 44625. 423–432.

[6] GPy. since 2012. GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy.

[7] Sandeep KS Gupta, Rose Robin Gilbert, Ayan Banerjee, Zahra Abbasi, Tridib Mukherjee, and Georgios Varsamopoulos. 2011. Gdcsim: A tool for analyzing green data center design and resource management techniques. In *2011 International Green Computing Conference and Workshops*. IEEE, 1–8.

[8] Xu Han, Wei Tian, Jim VanGilder, Wangda Zuo, and Cary Faulkner. 2021. An open source fast fluid dynamics model for data center thermal management. *Energy and Buildings* 230 (2021), 110599.

[9] Hrvoje Jasak, Aleksandar Jemcov, Zeljko Tukovic, et al. 2007. OpenFOAM: A C++ library for complex physics simulations. In *International workshop on coupled methods in numerical dynamics*, Vol. 1000. IUC Dubrovnik Croatia, 1–20.

[10] David J Lettieri, Michael M Toulouse, Cullen E Bash, Amip J Shah, and Van P Carey. 2013. Computational and experimental validation of a vortex-superposition-based buoyancy approximation for the COMPACT code in data centers. *Journal of Electronic Packaging* 135, 3 (2013), 030903.

[11] Lei Li, Chieh-Jan Mike Liang, Jie Liu, Suman Nath, Andreas Terzis, and Christos Faloutsos. 2011. Thermocast: a cyber-physical forecasting model for datacenters. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1370–1378.

[12] Zachary M Pardey, James W VanGilder, Christopher M Healey, and David W Plamondon. 2015. Creating a calibrated CFD model of a midsize data center. In *International Electronic Packaging Technical Conference and Exhibition*, Vol. 56888. American Society of Mechanical Engineers, V001T09A029.

[13] Tu-Hoa Pham, Giovanni De Magistris, and Ryuki Tachibana. 2018. Optlayer-practical constrained optimization for deep reinforcement learning in the real world. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6236–6243.

[14] Amir Radmehr, Brendan Noll, John Fitzpatrick, and Kailash Karki. 2013. CFD modeling of an existing raised-floor data center. In *29th IEEE Semiconductor Thermal Measurement and Management Symposium*. IEEE, 39–44.

[15] Emad Samadiani and Yogendra Joshi. 2010. Multi-parameter model reduction in multi-scale convective systems. *International Journal of Heat and Mass Transfer* 53, 9-10 (2010), 2193–2205.

[16] Emad Samadiani and Yogendra Joshi. 2010. Proper orthogonal decomposition for reduced order thermal modeling of air cooled data centers. *Journal of heat transfer* 132, 7 (2010).

[17] Emad Samadiani, Yogendra Joshi, Hendrik Hamann, Madhusudan K Iyengar, Steven Kamalsy, and James Lacey. 2012. Reduced order thermal modeling of data centers via distributed sensor data. *Journal of heat transfer* 134, 4 (2012).

[18] Mike Shafto, Mike Conroy, Rich Doyle, Ed Glaessgen, Chris Kemp, Jacqueline LeMoigne, and Lui Wang. 2012. Modeling, simulation, information technology & processing roadmap. *National Aeronautics and Space Administration* 32, 2012 (2012), 1–38.

[19] Umesh Singh, Amarendra Singh, S Parvez, and Anand Sivasubramaniam. 2010. CFD-Based Operational Thermal Efficiency Improvement of a Production Data Center.. In *SustainIT*.

[20] Lawrence Sirovich. 1987. Turbulence and the dynamics of coherent structures. I. Coherent structures. *Quarterly of applied mathematics* 45, 3 (1987), 561–571.

[21] Kunihiko Taira, Steven L Brunton, Scott TM Dawson, Clarence W Rowley, Tim Colonius, Beverley J McKeon, Oliver T Schmidt, Stanislav Gordeyev, Vassilios Theofilis, and Lawrence S Ukeiley. 2017. Modal analysis of fluid flows: An overview. *Aiaa Journal* 55, 12 (2017), 4013–4041.

[22] Wei Tian, Thomas Alonso Sevilla, Dan Li, Wangda Zuo, and Michael Wetter. 2018. Fast and self-learning indoor airflow simulation based on in situ adaptive tabulation. *Journal of Building Performance Simulation* 11, 1 (2018), 99–112.

[23] Ruihang Wang, Xin Zhou, Linsen Dong, Yonggang Wen, Rui Tan, Li Chen, Guan Wang, and Feng Zeng. 2020. Kalibre: Knowledge-based Neural Surrogate Model Calibration for Data Center Digital Twins. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 200–209.

[24] Christopher K Williams and Carl Edward Rasmussen. 2006. *Gaussian processes for machine learning*. Vol. 2. MIT press Cambridge, MA.

[25] Wangda Zuo and Q Chen. 2007. Validation of fast fluid dynamics for room airflow. *IBPSA Building Simulation* 2007 (2007).