THE OPTIMISED OPERATION OF COOLING DEVICES IN DATA CENTRES:

A VAR-PSO-BASED MPC SCHEME

Thesis submitted for the degree of Doctor of Philosophy By

Yang Hai

April 2023

Brunel Business School

Brunel University

Supervised by

Professor Habin Lee

PUBLICATIONS

Journals

- Y.Hai, P.Zhao, A comparison research on the financing strategies based on Garch-VAR models. Value Industries. 2019.
- -Lee.Habin, Choi.Youngseok, Van Nguyen.Truong, Hai.Yang, Kim.Junchul, Bahja.Mohammed, Hocaoğlu.Hakan., “COVID-19 Led Virtualization: Green Data centre for Information Systems Research,” Inf. Syst. Manag., vol. 37, no. 4, pp. 272–276, 2020.

Conferences

- Y.Hai. T.Nguyen,Y.Choi.M,Bahja, H.Lee. Operational Management of Data Centres Energy Efficiency by Dynamic Optimisation- Based on a Vector Autoregressive Model-Reinforcement Learning (VAR-RL) Approach, CERC Conference Proceedings, 2020.09
- Poster and PPT: A MPC Control Strategy for Data Centre Energy Consumption Optimisation-Based on a VAR-PSO approach, Ecocity 2022 conference

Working papers

- A Dynamic Cooling Approach to Energy Efficient Data Centres: A Model Predictive Control Framework, European Journal of Operational Research

ABSTRACT

Data centres (DCs) are the most significant energy consumers globally, where IT and cooling devices account for approximately 45% and 55% of their total energy consumption, respectively. Despite the extensive research conducted on reducing the energy consumption of IT devices, studies focusing on the reduction of energy consumption of cooling devices in DCs are relatively limited. Furthermore, there is a lack of research on the optimal utilisation of existing cooling devices to minimise their energy consumption. In this study, a Model Predictive Control (MPC) framework, in which a Vector Autoregressive Model (VAR) and Particle Swarm Optimisation (PSO) are integrated, is proposed to optimise the energy consumption in DCs by controlling the temperature setpoints of air conditioners (ACs).The VAR model is employed to capture the causal-effect relationships among the system variables, which affect the temperature changes in DC rooms, and then used to predict future temperature parameters over time. The PSO algorithm is utilised to find the optimal temperature setpoints combinations based on the future temperature changes provided by the VAR model. The Model Predictive Control (MPC) framework controls the VAR model and the PSO optimisation, continuously evaluating the optimised operation of air conditioners (ACs) and adjusting the VAR model if any deviation is detected to ensure the energy efficiency of DCs falls within the predefined range. Through this approach, the optimisation problem can be solved dynamically, taking future performance into considerations, and proactively avoiding potential issues. The feasibility of the proposed MPC framework has been tested in a national DC room in Thailand. Moreover, the effectiveness of the framework under various operating

scenarios was validated through a Computational Fluid Dynamic (CFD) simulated environment. The results of the field experiment demonstrate that the proposed MPC framework is effective in reducing energy consumption in the DC room, achieving a 32.5% reduction compared to existing cooling practices that utilise fixed AC setpoints during operation. Additionally, the simulation results illustrate a high adaptability of the proposed approach to changing conditions. This study makes significant contributions at the intersection of theoretical, methodological, and practical domains. In terms of theoretical contributions, this study challenges prevailing paradigms in DCs' energy optimisation. By emphasizing managerial solutions over traditional hardware modifications, our approach offered a novel perspective on effective energy optimisation strategies. Methodologically, our research introduces a dynamic framework (MPC) integrating predictive modelling (VAR) and optimisation algorithms (PSO). The integration of VAR, PSO, and MPC approaches not only leverages their respective strengths but also compensates for their individual limitations, maximising the synergistic potential. Our method also provides a flexible and adaptive solution by autonomously adjusting to changing system states. This adaptive quality bridges the gap between theory and practical application, differentiating our approach significantly from conventional practices in comparable industries. Practically, the proposed approach has been proved effective in controlling the temperatures in the DC room, achieving notable energy savings. Furthermore, the experiment demonstrates that the proposed MPC framework responds to workload changes within a reasonable timeframe, indicating its' real-time adaptability.

Keywords: Data centre, Energy consumption, Vector Autoregressive Model, Model predictive control, Particle Swarm Optimisation

ACKNOWLEDGEMENTS

I would like to express my great gratitude to my supervisors, Prof.Habin Lee and Dr Youngsoek Choi for their patient guidance and useful instructions on my writing and research skills. I would like to acknowledge the panel board who provide precious comments and suggestions for my all progress reviews. And I would like to thank my family and all the people who support me during my study, thanks Dr Tawan Boonkongwathana for his advice and assistance in proofreading, as well as for sharing valuable experience from his PhD study. Finally, special thanks should be given to our beloved P' Nam, you are our angel and will be memorised forever, we love you.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

NOTATIONS

T: set of time points, where $T = [0,...,T_{max}] \cap$ N.

P: set of lagged terms, where $P = [0,...,P_{max}] \cap$ N.

$x_t \in$ R$^{nx\times1}$: state vector at time $t \in T$, where n_x represents the dimensionality (number of components in x_t).

$u_t \in$ R$^{nu\times1}$: vector of control actions at time $t \in T$, where n_u represents the dimensionality (number of components in u_t).

$x^{ref}{}_t \in$ R$^{nr\times1}$: reference state vector at time $t \in T$, where n_r represents the dimensionality (number of components in $x^{ref}{}_t$).

$V = [1,...,V_{max}] \cap$ N: set of ACs units.

$J = [1,...,J_{max}] \cap$ N: set of the server racks.

$S = [1,...,S_{max}] \cap$ N: set of ceiling temperature sensors.

$Y_t = [y_{1t},...,y_{dt}]$: set of endogenous variables of the VAR model at time $t \in T$, where $d \in$ N denote the number of the endogenous variables.

$Z_t = [z_{1t},z_{mt}]$: set of exogenous variables of the VAR model at time $t \in T$, where $m \in$ N denotes the number of the exogenous variables.

$T_{v,t}^{cool}$: the temperature supply of the v^{th}AC unit at time $t \in T$, where $v \in V$.

$P_{v,t}^{cool}$: the power consumption for the v^{th}AC unit at time $t \in T$, where $v \in V$.

$P_{j,t}^{comp}$: the computational power for the j^{th} server rack at time $t \in T$, where $j \in J$.

$T_t^{outflow}$: the temperature at the airway point nearest to the AC units at time $t \in T$.

$WL_{j,t}$: the estimated DC workload (CPU usage) of the j^{th} server rack at time $t \in T$, $j \in J$.

WL_t: the average workload of the DC room at time $t \in T$.

$HG_{j,t}$: the heat generated by the j^{th} server rack at time $t \in T$.

$P_{j,t}^{Comp}$: the computational power for the j^{th} server rack at time $t \in T$.

$T_{s,t}^{ce}$: the temperature from the s^{th} ceiling sensor at time $t \in T$.

$T_t^{ambient}$: the server room ambient temperature at time $t \in T$.

$T_t^{outdoor}$: the outdoor temperature from the sensors on the outdoor chiller at time $t \in T$.

n: number of ACs' temperature setpoints combinations.

A_i: $D \times D$ coefficient matrices of Y, where $i = 1,...,P$.

B: $D \times M$ coefficient matrix of potentially deterministic regressors (exogenous variables).

ε_t: a D dimensional white noise process in the VAR model

Abbreviations

AC	Air Conditioner
AIC	Akaike information criterion
ANN	Artificial Neural Networks
API	Application Programming Interface
AR	Autoregressive model
ARIMA	Autoregressive Integrated Moving Average model
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CCM	Causal Chain Model
CFD	Computational Fluide Dynamic
CPU	Central Processing Unit
DC	Data Centre
DCR	Data Centre Review
DSS	Decision Making System
GA	Genetic Algorithm
GPR	Process Regression
IT	Information Technology

LQG	Linear Quadratic Gaunssian Control
MILP	Mix Integer Linear Programming
MIP	Mixed-integer Programming
MPC	Model Predictive Control
MPC	Model Predictive Control
NN	Neural Networks
OLS	Ordinary Least Square
PDU	Power Distribution Unite
PID	Proportional Integral Derivative
PSO	Particle Swarm Optimisation
PUE	Power Usage Effectiveness
SC	Schwarz Criterion
SI	Swarm Intelligence
SVR	Support Vector Regression
TS	Tabu Search
TSP	Traveling Salesman Problem
UPS	Uninterruptible Power Supply
VAR	Vector Autoregressive model
VMA	Vector Moving Average model
VM	Virtual Machine
WL	Workload
OT	Outdoor temperature

CHAPTER 1

INTRODUCTION

This chapter contains a brief introduction to the current state of energy consumption in global data centres (DCs) and an overview of the existing DC energy-saving operations. The rationale behind the need for an efficient and effective operational solution to the DCs will be presented. Additionally, the discussion covers the motivation, objectives, and significance of this study. The chapter closes with an outline of the dissertation.

With the increasing demands of digital computing services, Data Centres (DCs) have become one of the largest energy consumers. It was reported that DCs consume 100 times more energy than conventional office buildings (Jiao et al. 2017). Significant efforts have been made into improving the energy efficiency of DCs. However, such studies were mainly focused on the design of the infrastructures of the DCs that require plenty of resources and time by replacing their current facilities. This study aims at developing managerial strategies for using the existing cooling devices in DC rooms to achieve energy efficiency through active interventions.

Two major energy consumers in a DC room are cooling devices and IT devices, where cooling devices occupy more than 55% of the total power consumption (Zhang et al. 2017). There is a recognised trade-off dynamic between the energy consumption of cooling devices and IT devices within a DC. This trade-off stems from the fact that maintaining lower temperatures in the DC leads to an increase in the energy consumed by cooling devices. However, in the meantime this decrease in temperature enhances the efficiency of IT devices, resulting in reduced energy consumption on their part. Conversely, the decision to maintain higher temperatures in the DC results in a reduction in the energy consumption of cooling devices. Nevertheless, this reduction in cooling energy would potentially compromised computing efficiency and leads to a subsequent rise in the energy usage of IT devices. This delicate balance underscores the intricate relationship between temperature management, cooling systems, and the energy efficiency of IT infrastructure within a DC environment. The traditional operational strategy used by DC managers to save energy consumption of cooling devices is turning off a certain number of ACs to save energy in the winter season. However, a limited number of studies provide more sophisticated strategies to make optimal use of ACs. A detailed summary of the existing solutions can be found in Chapter 2. Such a research gap motivates this study.

Empirical pieces of evidence showed that increasing the AC setpoint by a single degree could result in 4-5% energy cost savings; and increasing the setpoint by 10 degrees, which is also a realistic number, could result in savings of over 40% (Glenn & Amherst 2020). Although this aspect is simple, it is hardly the case under the situation considering the complex nature of DC assets. According to the ASHRAE, (a US institute that provides DCs with guidelines for energy saving) thermal guideline from 2004 to 2016, the recommended thermal temperature has been changed from 20-25 degrees to a newer and wider recommended range of between 18-27 degrees from 2008 due to the advances in technologies for IT servers (ASHRAE 2014). To achieve an optimal envelope for a plant, it is essential to consider a trade-off between three crucial factors: high reliability, maximum energy efficiency, and permissibility limits within which the plant can function effectively. Finding the right balance between these factors is vital for ensuring the efficient and reliable operation of the plant. This is done through tests and is not a statement of reliability. In the current era, there is a larger focus on saving energy costs under the premise of ensuring the reasonable operation of the equipment instead of only looking at the reliability of the equipment. In practice, even referring to the thermal guidelines, DCs have the sense to maintain their temperature within the recommended range, however according to a survey, more than 90% of DCs maintain a constant temperature of approximately 24 degrees to safeguard the optimal functioning of IT devices. This practice stands in contrast to the recommended maximum working temperatures of 27 degrees by ASHRAE (2016), implying that DCs are potentially operating in an over-cooled state, contributing to inefficiencies in energy consumption. (Fernandes 2018). Increasing the AC setpoints blindly could jeopardize the health of the servers and other hardware, as existing hot spots may become even hotter and higher hot aisle temperature may activate server fans and offset efficiency gains (Ham et al. 2015). Therefore, a rigorous plan to optimise the ACs' temperature setpoints is critical to increase the energy efficiency of the DC.

From the optimisation perspective, the problem is challenging because of the following reasons. Firstly, the DC room environment involves multiple variables, such as all the energy consumption variables of the cooling components, temperature variables, power consumption variables, and the workload variables of the IT devices, as well as other variables. Therefore, the optimisation model needs to consider a large number of factors. Secondly, the interactions among the factors increase the complexity of the optimisation problem. Thirdly, DC rooms are prone to changes in room configurations, including the addition or removal of IT servers or server racks. An optimisation model formulated for a DC room configuration may not work anymore when the configuration is changed (E.Rob 2021). Therefore, the traditional optimisation approaches have limitations in the DC environment, and more dynamic and adaptive approaches are needed. In the fourth point, it is noteworthy that conventional optimisation methodologies frequently neglect the consideration of time lags in the interactions among variables. Specifically in the thermal contexts of DC rooms, where the time lags are prevalent. Consequently, it is crucial for the optimisation models that designed for the DCs environment conscientiously incorporate these time lags into their framework. The optimisation problem in this study aims to minimise the aggregate energy consumption by identifying the optimal combination of ACs temperature setpoints whilst maintaining the room temperature within permissible limits, thereby ensuring a conducive working environment for IT equipment. The integration of the Vector Autoregressive Model (VAR), and Particle Swarm Optimisation (PSO) with Model Predictive Control (MPC) represents a dynamic solution to the problem, leveraging one of the most advanced and powerful approaches in the field of process control. MPC model is a multivariate optimal control strategy that incorporates the predictive model and optimisation model in a process to optimise system objectives dynamically. There are many connections between MPC and optimisation. MPC and optimisation are closely linked, as MPC serves as an optimisation method for solving control problems. It provides optimised control actions for the controller by solving the optimisation problem. MPC consists of three main parts: (1) A system predictive model; (2) The prediction of the future movements of the related variables; (3) A control (decision-making) system. Following the process, a decision support system (DSS) was developed to optimise energy consumption in DCs using the MPC concept. The VAR model is utilised as the system predictive model in this context. It is a multivariate time series model that was initially introduced by Sergent and Sims in 1977, following the development of a dynamic structural Causal Chain Model (CCM) by Wold in 1964. The VAR model is a reduced form of a structural model in the Cowles tradition, as noted by Qin in 2011. VAR model in this system is used as a system analysis and forecasting model to capture the complex interactions among the number of meteorological variables and power plants and predict the future movements of the related variables after manipulating the control inputs, which are ACs temperature setpoints. Also, the VAR model will detect whether the constraints for IT server operations have exceeded the limits. Meanwhile, a PSO method solves the optimisation of objective functions over time in a receding horizon manner under the MPC scheme.

Compared to general optimisation approaches, the proposed MPC-based optimisation approach has the major advantage on the optimisation horizon that enables optimising the current state in time whilst taking future performance into account with the flexibility in terms of the capability to deal with multi-variables and handling interactions between them. This is mainly achieved by the following aspects: (1) VAR as a multivariate model is capable of mapping multi-variables in a complicated environment into one model, and the variables interact with each other bidirectional. (2) VAR serves as a predictive model that enables a forecast of future state conditions through the manipulation of control inputs (the ACs temperature setpoints). This capability facilitates the implementation of optimal inputs in advance, ensuring efficient energy consumption. (3) As a meta-heuristic optimisation solver, PSO can efficiently achieve the global optimal solution when compared to other types of optimisation solvers. PSO is particularly advantageous in handling multivariable optimisation problems and is easy to implement, further increasing its appeal as an optimisation tool. (4) MPC is a feedback receding horizon control system, that the output of the previous stage that will be used as the input to the next stage. It can predict future events and take control measures accordingly. The whole system can be run dynamically. (5) The MPC process allows multiple constraints on the control input and the states at the same time, this feature guarantees the optimisation system functions under multiple conditions.

The remaining paper will be organised as the following: Chapter 2 will provide reviews of the relevant literature on three aspects regarding the research gap on this topic and a detailed comparison of the existing approaches to solve the problem on a technical level. Chapter 3 discusses the typical processes involved in design science studies and highlights how this study aligns with the principles and methodologies of the design science field to address the identified problem. Chapter 4 provides an overview of the VAR-PSO-MPC approach, and the formulation of the optimisation problem will be presented. In Chapter 5, the focus is on the practical DC experiment and system implementation, which encompasses the VAR modelling and forecasting, Power Usage Effectiveness (PUE) optimisation, MPC control processes, as well as other supporting analyses necessary for the study. Furthermore, a comparison of performance between the proposed approach and the Genetic Algorithm is provided. Chapter 6 validates the real-world implementation of the VAR-PSO-based MPC scheme, while Chapter 7 delves into the use of simulation as a means to validate performance across various DC scenarios. Chapter 8 provides a summary of the research findings and discusses the advantages of this study compared to similar research. Additionally, it examines the future prospects of the technology, including potential challenges and opportunities for further research.

CHAPTER 2

LITERATURE REVIEW

In this chapter, a review of relevant previous research work is given on the existing energy-saving solutions of DCs. Three categories of the most popular energy-efficient engineering design techniques for DCs are introduced in this chapter's opening section, along with their limitations. This provides the methodological motivations for this study. Subsequently, the system identification models, optimisation approaches, and control theories are discussed in the context of industrial energy-saving operations. Detailed comparisons of similar approaches are also given and followed by a review of the relevant work in each field. Finally, the state of the arts of the proposed methodology in terms of the literature review and their relevance to the subject matter of this dissertation is stated at the end of this chapter.

2.1 The existing energy-saving solutions of DCs

Notably, although DCs has been used for years, their energy-efficient concern has not drawn much attention until recently, some research demonstrated a remarkable annual growth rate (20% to 33%) of global electricity consumption (Nadjahi *et al.* 2018). Such a result is inconsistent with the previous discussion about DC operators being more focused on the reliability of the devices instead of energy efficiency. Following the discovery of the unavoidable fact that DC has the biggest energy consumption worldwide, a concerted research effort has been put into the energy efficiency solutions of the DCs. This section will outline the most recent developments towards the solution to the problem, and the state of the art regarding the motivation of the proposed approach.

Current research on reducing energy consumption in DCs can be categorised into three main fields: cooling configuration design, IT-side renovations, and thermal management. The cooling configuration design mainly focuses on optimising the layout and efficiency of cooling systems to enhance heat dissipation and reduce overall energy usage in DCs. IT-Side renovations involve improvements and innovations in information technology infrastructure to enhance energy efficiency, such as more energy-efficient hardware and software solutions. Thermal management, as a more wild concept, compasses strategies for effectively managing and controlling the temperature within DCs, including advanced monitoring systems and adaptive thermal management techniques. In the upcoming discussion, the three primary fields will be reviewed individually and subsequently compared to the proposed solution presented in this dissertation.

(1) Cooling configuration design

The study by Cheung et al. (2018) realised that approximately 40% of the energy consumption goes into the cooling system and it is the major source of a high PUE in the DC, therefore a concerted research effort has been put into the development of cooling technologies. Traditionally, air cooling systems have played an important role in DC cooling technologies, innovations to improve the air-cooling energy efficiencies, such as containment of hot/cold aisles and variable fan speed has shown to reduce energy cost consumption significantly by up to 70% (Carbó *et al.* 2016a). Nevertheless, IT density doubles every two years according to Moore’s law, which indicates that traditional air cooling is insufficient to meet the heat generated by the increasing IT density. Following the inference by Moore, efforts have been made on novel cooling technologies, which focused on four categories as follows:

The free cooling technology aims at using the natural airside or water side economisers as well as the heat pipe air exchange technology to cool down the DC (Daraghmeh & Wang 2017).

The liquid cooling technology is efficient when the power density of the DC is high, there is plenty of evidence indicating that has the potential to solve many of the problems associated with air cooling systems, especially as computer density rise (Carbó *et al.* 2016b).

The two-phase flow technology utilizes a refrigerant that releases the heat generated by the server racks into the environment (Riofrío *et al.* 2016).

The building envelope technology such as building shell, fabric or enclosure materials and performances have also been wildly studied by the existing literature (Akeiber *et al.* 2016).

(2) IT-side renovation

According to New York Times, the average server usage rate is 6% to 12%, and 90% of the servers run at less than 5% utilisation (New York Times 2012). Efforts have been made to improve server performance, including the following:

Servers consolidation methods suggested integrating multi-redundant servers into fewer servers and sharing fewer power supplies (Verma et al. 2009).

Server virtualisation allows tasks to be done by multiple virtual servers instead of single physical servers to achieve an energy-saving purpose (Schulz 2011).

Storage consolidation is to manage the data storage to avoid the devices running with heavy storage that causes energy waste. It has been wildly suggested to adopt new alternatives to traditional disk storage as flash storage (Zhang *et al.* 2018).

Decommissioning the idle servers enables energy efficiency because energy savings from the server level result in approximately 1.9 times thus at the facility level due to reduced energy waste in the power infrastructure (PDU, UPS, building transformers) and reduced energy needed to cool the heat produced by the server (Emerson Network Power, 2012).

Upgrade energy-efficient power suppliers such as PDUs, PSUs, voltage regulators, processors as well as energy-efficient fans (Lintner *et al.* 2011).

Upgrade the CPU with a dynamic energy-saving function that is capable to run in an energy-saving mode according to the CPU usage (Lintner *et al.* 2011).

Upgrade the network transmission equipment to the newer generations as they pack greater throughput per unit of electricity (Wang *et al.* 2012).

Upgrade the constant fan speed cooling devices to variable fan speed, therefore adapting to the changes in DCs workload (Pöyhönen *et al.* 2021).

(3) Thermal management

Adapting the server racks to the hot/cold aisle layout as well as the containment and enclosure methods are the most commonly adopted approaches to saving energy in cooling (Niemann *et al.* 2013).

The optimal layouts of the cooling devices according to the thermal dynamics are also wildly studied by the literature. For example, shorten the distance between cooling devices and IT devices to minimise the cooling air waste on the way. (Stahl & Sullivan 2001),(Patel *et al.* 2002).

(4) Comparison of existing solutions

Although there are remarkable breakthroughs in the field of the DCs' hardware innovations, however, one of the common disadvantages of the existing approaches is the complexity and high cost of upgrading and replacing the existing infrastructure. Existing new technologies require additional time to verify, improve and mature before applying to the practical DCs. The high cost of replacing the existing facilities is also crucial for realising the energy-saving goal of the DCs through hardware renovations and replacement. The report from Forbes Technology Council (R. Danilak 2020) evaluated that it would be hardly seen some essential improvements in the above areas in the next five years since a benefit from the adoption of new energy-efficient techniques for DCs would be easily compromised by the cost of renewing the DC infrastructures, this essentially became one of the barriers to realising the facilities renovations. Therefore, this circumstance sparks our motivation to investigate managerial strategies instead of changing the existing facilities in DC to reduce energy costs.

Figure 2-1 shows a systematic review of more than 200 related studies from our previous work. shows that around 39% of the literature look at the IT facilities, 33% look at cooling facilities and 19% focus on thermal management in the DC. Most of the papers written were on observations and ideas from data centres and an analysis of them, rather than a movement towards a solution. There is a limited number of studies investigating cooling operations to save energy in DCs without changing any infrastructure. This highlighted the state of the art of our proposed approach to fill up the research gap in the

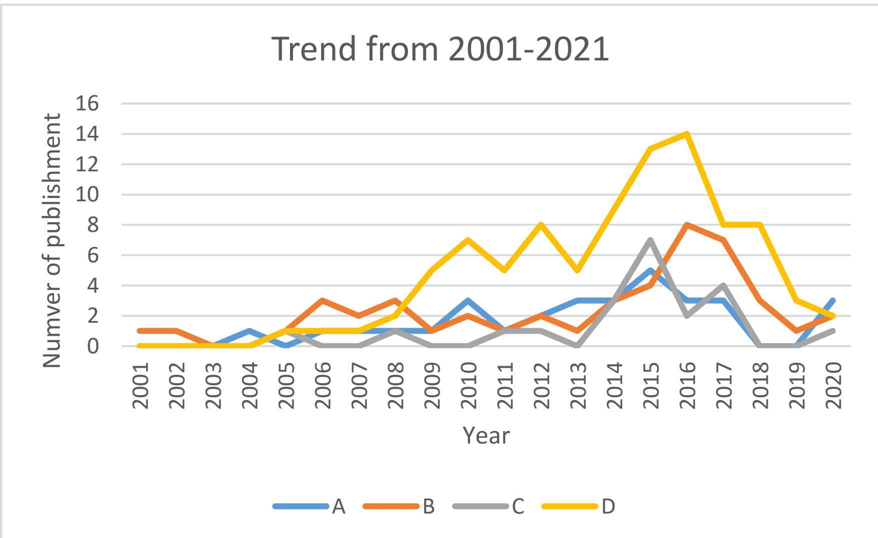


Figure 2-1 The trend of literature

existing literature.

A – Literature that investigates innovations in cooling facilities. **B** – Literature that includes information related to thermal management. **C** – Literature review papers. **D** – Literature that investigates innovations in IT facilities

Table 2-1 provides a comparative overview of current DC energy-saving solutions, as well as presenting the positioning of the present thesis within the existing literature.

Table 2-1 The comparative overview of DC energy-saving solutions

Field	Reference	Exemplary Approaches	Key Findings
Cooling configuration design	Daraghmeh and Wang (2017)	A free cooling technology	The integrated system of mechanical refrigeration and thermosiphon (ISMT) is considered an ideal solution for DCs. The improved ISMT consumed about 34.3 –36.9% less energy than traditional cooling systems.
	Carbó et al. (2016b)	Liquid cooling technology	The study analyses and quantifies the potential for heat reuse by performing extensive thermal characterisation and suggested that the use of liquid cooling systems in DCs can provide higher heat removal capacity.
	Riofrío et al. (2016)	The two-phase flow technology	Analysed the enhancements of two-phase flow cooling technologies including microchannel technology, plate-fin heat exchangers (PFHE),

			and spray cooling, also provide evidence of their efficiency.
	Akeiber et al. (2016)	The building envelope technology	Reviewed popular building envelop technologies and suggests Phase change materials (PCM) can be incorporated into the building envelope to increase its thermal mass and improve energy performance.
IT design	Verma et al. (2009)	Server consolidation	Adopted a pMapper power-aware application placement methodology and successfully reduced power consumption in server side through consolidation.
	Schulz (2011)	Server virtualization	This study revealed that the implementation of servers virtualisation facilitates the utilisation of a reduced number of servers, resulting in reduction in electricity consumption and waste heat. It also suggests that decommissioning the idle servers will lead to significant energy saving.
	Zhang et al. (2018)	Storage consolidation	Designed a Combining three (C3) system that consists of three modules: prediction, consolidation, and migration to achieve live migration and reallocation of the server's storage, therefore achieve energy reduction.

	Pöyhönen et al. (2021)	Server hardware innovations	By upgrade the constant fan speed cooling devices to variable fan speed that could adapt to the server's workload fluctuations, the energy consumed by the servers can be significantly reduced.
Thermal management	Niemann et.al (2013)	Optimal layout of servers	Energy reduction can be observed by adapting the server racks to the hot/cold aisle layout as well as the containment and enclosure methods.
	(Stahl & Sullivan 2001),(Patel et al. 2002)	Optimal layout of cooling devices	Shorten the distance between cooling devices and IT devices can minimise the cooling air waste during transmission.
Managerial strategy	This study	Optimise energy consumption by optimise the ACs temperature setpoints	Significant energy reduction of 32.5% without any hardware renovation or infrastructure changes.

2.2 Technical methodological review

This section will review the literature at a technical methodology level. As the proposed approach is the combination product of three fields: system identification, optimisation problem solving and the control strategy, similar approaches will be compared and evaluated respectively in each field. The current research gap, current challenges, fierce debates, as well as the state of the art within each field will be discussed on the technical level.

2.2.1 System-identification in modelling optimisation problems

System identification is a fundamental methodology for modelling complex system and subsequently optimising their behaviour. It's a process that involves extracting the insights of the systems' dynamics by observing the interactions between the system variables. The system identification process takes inputs represents the stimulus or control applied to the system, while the output signifies the system's corresponding response (Theodoridis and Kraemer). When transitioning to the realm of optimisation, the input and output relationship of the system identification model becomes instrumental. Optimisation aims to find the most efficient or optimal set of inputs that lead to desired outputs. In this context, the system identification model acts as a surrogate for the real-world system, facilitating the exploration of input configurations that yield optimal outcomes. By utilising the recognition gained from the system identification model, optimisation algorithms can navigate through the input space, seeking combinations that maximise or minimise a specified objective function(Biegler et al.).

Selecting the appropriate identification method is seamlessly aligns with the identification of the optimisation problem. Optimisation of energy consumption through the adjustment of AC setpoints is directly connected to combinatorial problems. In the context of air conditioning management, determining optimal setpoints involves discrete decision-making, presenting a challenge analogous to those explored in combinatorial optimisation studies (Boyd 2004). These types of problems have been extensively studied, ranging from Traveling Salesman Problem (Prabowo et al. 2018; Tadei et al. 2017), scheduling problems such as Flow-shop and Job-shop scheduling (Calude et al. 2005), 0-1 knapsack problems (Feng et al. 2017), Bin Packing problems (Song et al. 2018), and graph clustering problems (Bastkowski et al. 2016). Among those, the scheduling problem emerges as particularly relevant to our current exploration. Extensive research has explored energy-efficient scheduling, particularly in the context of thermal dynamics. The Empirical mathematical equations, grounded in first principles, serve as foundational elements for a wide range of commercial modelling software. (US.DOE.2014). However, although traditional mathematical modelling occupies an important place in optimisation modelling, several practical problems remain to be addressed. Studies have proved that traditional mathematical modelling takes up to 90% of the effort regarding the time and cost in the modelling process, particularly when the environment is multivariate and complicated. (Chemical Engineering Department King Saud University 2002). Due to the demanding need for an in-depth understanding of a system's nature. The challenge of traditional mathematical modelling arises from the complexity of the system or limited knowledge of the underlying physical laws and principles (Mäkilä & Partington 2003). Alternatively, data-driven model has emerged to be a shortcut to address the inherent challenges proposed by traditional mathematical modelling.

Perez, Baldea and Edgar (2019) adopted a data-driven approach to estimate the coefficient of a Reduced Order Model (ROM) as the objective function of the problem is to minimise the peak hours’ cooling load by pre-cooling control actions. The ROM model largely reduced the computational workload in real-time. This is the most related paper in the literature that studies the DCs cooling managerial strategy. However, it revealed several gaps and shortcomings. This research focused on shifting the peak hours' cooling load to the unoccupied time to pre-cool the house, therefore achieving the peak load reduction instead of managing the energy-saving operations all over time. Another major drawback to adopting the ROM model is that the model is inherited from the empirical studies of the building cooling environment, to ensure computational efficiency, the independent variables are fixed to be a few main impactable components, in other words, ignored the variables that less impactable to the cooling load. In specific cases, the ROMs are invalid because the empirical model only considered the generalised household thermal environment, however, not able to adapt to special cases. Most importantly, this model fails to take the mutual-affected relationships among the variables in the complex environment into account due to its simplicity. Although the energy-saving feasibility has been proved by the data from the individual household, however, because of the similar residential behaviours and the relatively simple household environment, the ROMs are insufficient to adapt to a more complex thermal environment, such as heavy-duty DCs. The proposed system identification VAR model aims to leverage data-driven models, at the same time to address the gaps mentioned above.

The complexity of the DC environment, which encompasses numerous variables, is widely recognised (Ferreira & Pernici, 2016). As such, there is a need for a multivariate model that can capture the interactions between these variables. The proposed VAR model is one of the most commonly used models in multivariate time series analysis. From a longitudinal perspective, the Vector Autoregression (VAR) model is an extension of the univariate time series Autoregressive (AR) model. From a horizontal viewpoint, the VAR model, together with its related models, such as Vector Moving Average (VMA) and Vector Autoregressive Moving Average (VARMA), employ linear relationships to describe a stationary system. Compared to the general time series estimation model and some machine learning techniques, VAR has various advantages. It strengthens by describing the interaction relationship among variables with fast, easy and efficient features. Athavale *et al.* (2019) compared the temperature prediction performance of four different types of Data-Driven Models (DDMs) including Artificial Neural Networks (ANN), Support Vector Regression (SVR), Gaussian Process Regression (GPR) as well as Proper Orthogonal Decomposition (POD) in a DC, the result demonstrated that only NN could handle multiple output points in one model. However, because of the unknown features of the system and multi-dimensional problems that need to be solved in one model, it requires a large volume of data to feed into the model and all these models are facing a similar difficulty, which is expensive to run in practice and relatively time-consuming. In contrast, VAR considered all the endogenous variables into a mutual-effected relationship and represents the system in a comprehensive matrix form, because of its linearity feature, the model is inexpensive and efficient to run. Moreover, as a time series model with the AR structure, the data can be substituted into the VAR model recursively to predict the future movements of the system variables. E.Mostafa (2015) compared the performances of several macroeconomic structural models to the benchmark models VAR and

Autoregressive Integrated Moving Average (ARIMA) models, the result demonstrated that the VAR model outperformed the other models regarding the forecasting accuracy and the amount of interpretable information contained in the prediction result. A psychological study done by Eason (2020) demonstrated that the VAR model had a good performance in analysing the bidirectional relationship between the objectives subject to the selection of variables and the size of the sample data. Compared to the other nonlinear models, VAR has advantages from the linearity property, it simplified the estimation procedure and, most importantly, gives an easily interpreted output on a modular level in terms of the parameter weights (Friedman *et al.* 2000).

Table 2-2 presents a comparison between traditional mathematical methodologies and data-driven approaches. Additionally, within the data-driven model's scope, it contrasts the VAR model with various system identification models commonly referenced in the literature, emphasizing their applicability to the current study.

Table 2-2 The comparison of VAR and other system identification models

Category	**Model**	**Advantages**	**Disadvantages**	**Applicability**
Traditional mathematical approach	Queuing Models; LP,IP,MIP; Thermal models; Makov models; Game theory models	(1) Provides a clear optimization framework for resource allocation. (2) Allows for the incorporation of constraints and objective functions in a systematic manner. (3) Well-established algorithms for solving LP and IP problems efficiently.	(1) Computational Expensive (2) Requires deep understanding of the systems physics. (3) Lack of flexibility to adapt to the changes in environment	Consider the complexity of the DC environment, and the requirement of flexibility, adaptability of the model itself, traditional mathematical approach is not appropriate.
Data-driven approaches	VAR	(4) Linearity properties include efficient implementation and computational saving, more	(4) The performance of forecasting may be affected by the selection of variables and the size	(1) DC is a complex environment in that a multivariate model is required.

		interpretive compared to non-parametric models. (5) Parameters are easy to estimate: Under stationary conditions, the parameter estimation of VAR is consistent with the ordinary least square estimation (OLS). (6) Has many statistical properties: Allowing interval estimation, residual analysis and model diagnosis, etc. (7) The Time series models have advantages in forecasting. (8) Allows involving multi-variables. (9) Endogenous, exogenous assumption makes the interrelationships between variables easy to interpret. (10) Flexible, easily increase or reduce variables.	of the sample data. (5) Compared to other structural models, VAR parameters may have less theoretical meaning.	(2) Sample data is enough to estimate a VAR model and its efficiency makes real-time tracking and control possible. (3) It is an appropriate approach for the DC environment because it doesn't need a thorough understanding as well as time and effort to study the systems physics.
	ARIMA	(1) Linearity. Easy to implement; (2) Only include one endogenous variable and its past values. The structure is simple; (3) Can be used for forecasting purposes; (4) Seasonality adapted.	Univariate model. Not apply to the complex environment.	As a univariate model, ARIMA is not suitable because DC is a complex system with multi-variables.
	Classification models-SVM	(1) Able to solve high-dimensional problems, and large feature spaces. (2) Able to solve the problem with small samples. (3) Able to handle the interaction of nonlinear features. (4) No local minimum problem; (relative to ANN and	(1) The efficiency is relatively low When the sample size is large. (2) There is no universal solution to nonlinear problems, and sometimes it is difficult to find a suitable kernel function. (3) Not sufficient in interpreting high-	Not suitable because : (1) There is a large size of sample data in DC. (2) DC has multi-variables the binary classification is not appropriate. (3) Real-time estimation of the DC environment has the

		other algorithms). (5) No need to rely on the entire data. Strong generalization ability;	dimensional mapping of the kernel function, especially the radial basis function. (11) Conventional SVM only supports binary classification. Sensitive to missing data.	chance to receive missing data in the sample that SVM is not satisfied to handle.
	ANN	(1) Has the ability to learn and model nonlinear complex relationships. (2) ANN can be generalized. After learning from the initial historical data, it can also infer the unknown relationship of unknown data. (3) ANN does not impose any restrictions on the input variables (distribution). Many studies show that ANN performs well in modelling data with heteroscedasticity because it can learn hidden relationships in the data without imposing any fixed relationships. This is very wildly used in financial time series forecasting with very large data fluctuations.	(1) Poor interpretative. The most likely known disadvantage of neural networks is their "black box" nature (unknown law about the outputs). (2) Time costly. ANN takes a long time to train. (3) Compared with traditional machine learning algorithms, neural networks usually require more data, at least millions of labelled samples. However, many machine learning problems can be solved well with less data. (4) Computational expensive.	Not suitable because: (1) DC environment is a serious environment that needs an interpretative and trust model as a system model. (2) The DC needs to make a fast and efficient reaction to the energy inefficiency, computational-expensive would be a barrier to this purpose. Also, the ANN itself will consume more energy.
	Deep learning	(1) Compared with machine learning, deep learning is more effective on complex problems. (2) More potential depends on the sample size.	(1) Computational expensive. (2) Required high-performed hard-core. (3) “Black-box” property. Less interpretative. (4) Designing the model is complicated.	Similar to ANN, deep learning is the advanced version of the general neuron network which has more layers, this will increase the level of complicity and increase computational and human effort, moreover, time-consuming.

2.2.2 Review of the optimisation solvers

As the magnitude of data increases and hardware capabilities improve, optimisation issues tend to be tied to more scenarios, and the dimensions of the algorithms grow proportionately. One way to solve such high-dimensional data is to impose certain structural constraints on the problem from the perspective of parameter estimation, which is often non-convex. Correspondingly, the objective functions of such optimisation problems are also non-convex. Non-convex objective functions and constraints can model the problems more accurately, however tackling such problems might be difficult. Compared to convex optimisation problems, non-convex optimisation issues are more difficult to solve. Solving the objective function by traditional mathematical approach is mostly NP-hard, and addressing approximation solutions is also likewise NP-hard (Jain & Kar 2017). R.Urbanucci (Urbanucci 2018) categorized the optimisation solutions for energy polygeneration systems into three types, including optimum synthesis, design, and operations, in which MILP (Mixed Integer Linear Programming) is the approach for operations with the main advantage in finding the global optimal solution to the problem and easy to be solved by many commercial solvers. In practice, solving optimisation problems in high dimensions (large candidate solutions) can be difficult because it is time consume and computationally expensive during calculation by mathematical evolution algorithms (Tomassetti & Cagnina 2013) or nonlinear algorithms (Du & Chen 2000). D.Steen (2015) derived a thermal storage model for energy-saving by MILP and confirmed that because of the existence of the endogenous problem in the thermal environment, the temperature storage cannot be tracked by MILP for the whole timeline horizon but only a single time step. To overcome these difficulties, the proposed approach integrates endogenous modelling by a time series model VAR and solves the VAR objective function with a PSO solver.

In recent years, there has been significant attention given to addressing the challenge of solving high-dimensional non-convex optimisation problems. The most widely adopted approaches are gradient descent, momentum, and heuristic. Swarm Intelligence (SI) is a popular heuristic approach used to solve high-dimensional non-convex optimisation problems. It was introduced in 1989 by G. Beni and J. Wang (1993) and inspired by the flocking behaviour of birds in biology. Following the emergence of SI, Kennedy and Eberhart (1997) developed a heuristic approach named Particle Swarm Optimisation (PSO) in the mid-1990s. PSO is a computational method that can improve the optimisation of candidate solutions through iteration. It obtains a set of candidate solutions according to the mathematical formulations of the position and velocity of the particles, and the movements of particles in the search space give potential solutions to the problem. The movement of each particle is not only affected by its local optimal position, but the global optimal solution will also guide these particles to the optimal position. PSO update the optimal information by the joint feedback from each particle (cognitive) and its neighbour particle (social) that guaranteed the high probability and efficiency concerning finding the global optimal. Due to its various benefits including robustness, performance, and simplicity, PSO has been wildly accepted. In general terms, the PSO algorithm is mainly used for solving contentious variable problems. However, because of the good convergence and adaptability of PSO, efforts have been made to adapt PSO to solve the discrete optimisation problem as well. Kennedy & Eberhart (1997) defined the first discrete binary version of the PSO algorithm. The particles are encoded using binary strings. By using the sigmoid function, the speed is limited to the interval [0, 1] and interpreted as a "change in probability". Yang et al. (2019) extended this method to quantum space. Similarly, Pang et al. (2004) used a fuzzy matrix to represent the position and velocity of the particle, redefined the operator of the PSO algorithm, and applied it to the solution of the Traveling Salesman Problem (TSP) problem. G.Pampara (2006) combined the PSO algorithm with the angle modulation technology in signal processing to reduce the dimensionality of the high-dimensional binary problem therefore the optimisation problem can be solved. Afshinmanesh *et al.* (2005) redefined the addition and multiplication in the discrete PSO algorithm and uses the negative selection in the artificial immune system to achieve the speed limit. Hu (Hu *et al.* 2003) proposed an improved PSO algorithm to deal with the permutation problem, preliminary research on the n-queen problem shows that the improved PSO algorithm is very promising in solving the constraint satisfaction problem. Parsopoulos & Vrahatis (2000) used standard functions as an example to test the ability of the PSO algorithm to solve integer programming problems. A.Salman (Salman *et al.* 2002) abstracted the task assignment problem as an integer programming model and proposes a solution based on the PSO algorithm. It was observed that PSO requires less computational effort and has fewer parameters to adjust compared with other stochastic algorithms such as genetic algorithms (Khare & Rangnekar 2013). Mahor

(Mahor *et al.* 2009) reviewed the general optimisation approaches and the result illustrated that the heuristic approaches such as genetic algorithms, simulated annealing, ant colony optimisation, neuron network, Tabu search, and PSO performs better than general conventional optimisation approaches such as mathematical programming, and among them, PSO outperforms the others among those and provided the best result to the question. Table 2-3 compared and summarised the advantages and disadvantages of the penitential approach to the problem from the literature, and the applicability of chosen PSO to solve the proposed optimisation problem.

Table 2-3 The comparison of PSO and other common optimisation approaches

Category	Model	Advantages	Disadvantages	Applicability
Conventional optimisation approach	Mathematical programming	1) Clear structure, good interpretation of the relationship of the variables 2) Robustness and the fixed structure make the model perform static and robust. 3) Easy to implement. Fast speed 4) Efficient during calculation. 5) Simple and compact approximation to complex decision-making.	Requires a deep understanding of the nature of the environment. Construction of the model can be time and human-effort-consuming. Can be complicated when the number of variables increases. Solving such a complicated algorithm is difficult or even cannot be solved. Re-estimation of the system would be problematic. Hard to adapt to dynamic changes.	The optimisation formulation will take advantage of mathematical programming, using statistical parameters to form the objective function and constraints. It will also consider the dynamic revolution of the environment. But we solve it with the heuristic approach to avoid the complexity of finding the mathematical solution.
Heuristic	Tabu Search (TS)	1) Fast, efficient to converge to the globally optimal. 2) Easy to implement.	Too greedy to the local or neighbour spaces, then ignored the possibility of other spaces. Easy to stick in local optimal.	Too strict and greedy to the optimal solution, however, in the DC environment we expect to find the global optimal with a fast converge speed.
	Genetic Algorithm (GA)	The optimised function is not required to be differentiable, continuous, etc. The search process starts from a set of solutions to	Compared to PSO, it has the shortcoming that it does not have any memory of previous attempts. All the information is shared between individual participants	Compared to PSO, GA is less efficient with the converging speed. In real-time optimisation practice, PSO is preferred. Programming is more complicated. Once the

		the problem, rather than from a single individual. It has an implicit parallel search feature, thereby reducing the possibility of falling into a local minimum.	in GA, however, PSO only shares optimal solutions among individuals, resulting in GA converging slower than PSO to the optimal. Programming is more complicated. Encoding and decoding are required before and after optimisation is done. Crossover and mutation operation is problematic.	model needs to be adjusted, this increased the human effort. PSO is preferred.
	PSO	Has the advantage to memorize the previous attempts. The optimised function is not required to be differentiable, continuous, etc. Fast convergence speed. simple, easy to implement. The search process starts from a set of solutions to the problem, rather than from a single individual. It has an implicit parallel search feature, thereby reducing the possibility of falling into a local minimum. The input of decision variables can be multi-values	Although the PSO algorithm provides the possibility of global search, it does not guarantee convergence to the global optimum. At present, PSO has a little rigorous theoretical basis. It simulated the biological searching behaviour, but it does not explain the reason why it is effective and the scope of its application is relatively unclear.	The PSO algorithm is derived from the Complex Adaptive System (CAS). The subject is active and interacts with the environment and other subjects. Finally, the entire system may also be affected by some random factors. The above features are all consistent with the assumptions in the DC. Suitable to solve real-time optimisation problem which requires convergence speed. ACs' temperature setpoint combinations are in a large searching space, and PSO is strong in

		but not only one.		searching among high dimensions space.

2.2.3 Review of the system control strategies

In the field of online control, MPC is a widely used feedback control algorithm and has been recognised and applied in many industries. As a control approach, it comes from industry practice rather than academia. Track back to the early 1960s, only simple input/output models were applied for system control until Kalman (1960) raised the notion that the states are the “internal memory” of the system and not always measurable. This laid the foundation of Linear Quadratic Gaussian Control (LQG). However, LQG does not take the constraints into account and only can be applied to linear models. The problem has been solved after the receding horizon concept emerged. A model predictive heuristic control (MPHC) approach was introduced by Richalet in 1978 that has all the properties of MPC except the requirement of optimal controlling. In contrast to MPC optimisation, MPHC determines future actions by iteratively operating the system until the constraints are satisfied. While it only guarantees that the system will not violate the constraints, it does not necessarily ensure optimality. Nonetheless, it paved the way for the subsequent development of MPC, as noted by Richalet et al. (1978). Dynamic matrix control (DMC) was the first attempt at MPHC with the computation of optimal control variables, and it was initially applied by the Shell Oil Company in the 1960s. This approach has been reported to outperform the widely used Proportional-Integral-Derivative (PID) control in Shell Oil's production line.

Honeywell had confirmation of its convergence/stability in an internal study by the late 1970s/early 1980s. Beginning in the late 1980s, academic work on MPC demonstrated ways to lower the computing cost of MPC in critical issues. With the improvement of computational capabilities, MPC has been applied exclusively in industries (Mayne *et al.* 2000a) (B . Lee and L . Markus 2021). MPC has been introduced as a control technique for obtaining feedback from the predicted future state and optimising the active intervention to the current state, from the rapid control loop, the state can be optimised by the optimal control input at each time step. In this way, the whole system can be segmented into short-period linear processes and optimisation can be applied to each time horizon (Rawlings 2000). From the 1970s to nowadays, MPC has been developed from a heuristic control algorithm into a new branch of enriching theoretical and practical content. MPC is aimed at controlling problems with optimisation requirements, there is a variety of literature that has verified its successful control in complex industrial processes has been fully demonstrated and proves that MPC has a great potential for handling complex constrained optimal control problems (Qin & Badgwell 2003) (Mayne *et al.* 2000a). It has been realised by embedding the system identification model and prediction model in the MPC process. In recent years, MPC has been wildly used in many fields, such as advanced manufacturing, energy, environment, aerospace, and medical care. Literature about MPC covered the content of the supply chain for semiconductor production management (Wang *et al.* 2007), high-pressure composite processing in material manufacturing (Dufour *et al.* 2004), building energy-saving control (Salsbury *et al.* 2013), flight and satellite attitude control (Silani & Lovera 2005). A breakthrough in MPC is applied in the vehicle control conducted by Hassan (2016), an electronic vehicle speed control using MPC has been designed with the time delay based on a fuzzy model, and the simulation demonstrated a good control result of leading the electronic car to the motor with its speed setpoints) (Hassan Khooban *et al.* 2016). Since MPC is a flexible approach that has been already approved by the founder that it can be adapted to over 90% of the control problems, in recent years, MPC has been also considered to apply to the power system as well as the DCs. Izawa (Izawa & Fripp 2018) adopted an MPC control for the air conditioner energy consumption in the commercial building and take weather, electricity price, and occupancy into account. The result demonstrated the proposed approach can save 15% of the energy consumed by the air conditioner compared to other optimisation approaches adopted. Also, Google has reported that MPC control has significantly reduced the cooling cost by 9% in one of their DCs (Schwenzer *et al.* 2021). Yu et al. (2017) conducted a full set of benchmark approaches, including MPC to control the temperature in a small-scale building environment, and the result showed that MPC outperforms other approaches, including the most wildly-used thermostat technique, energy consumption in the DC has been convincingly reduced by 43% under MPC control strategy. Since the majority of the problems mentioned above are multi-variable and require constraints, MPC became a natural fit based on its strength in handling multi-variable issues and its' ability to simplify the complex system, and it has aroused the attention in the field of DC cooling control.

Technically, based on the principles of working, the control strategies can be divided into three categories: Classical controllers, predictive controllers and repetitive controllers (Schwenzer *et al.* 2021). As one of the predictive control strategies, MPC has advantages in many aspects compared to conventional control strategies, also among other predictive control strategies:

(1) The classical controllers and repetitive controllers, such as the PID controller, bang-bang controller or state controllers consider only the past and current system behaviours and are reactive to deviations (Ang *et al.* 2005) (Hillerström & Walgama 1996). However, MPC takes the future system behaviours into account, and the foresight of the prediction horizons will be considered to avoid the future problem in advance by minimising the deviation between the prediction and the reference (Rawlings 2000) (Schwenzer et al. 2021).

(2) Conventional controllers frequently encounter limitations as they primarily depend on accurately tracking the dynamics and trajectories of the system, which is considered the most challenging aspect of the control process. In contrast, MPC is presented as a more versatile solution that can effectively address a wide range of problems. Notably, MPC is noted for its capability to handle challenges even in situations where there is limited understanding of the system's nature or when there is a lack of confidence in the feasibility of traditional control approaches. Furthermore, MPC has the benefit from it relies on the optimisation model that fits any scenario, instead of formulating the tedious system control law, it determines the control law automatically

by a model-based optimisation. This main advantage makes MPC flexible to adapt to different industries and favourable to the engineering community (Lee *et al.* 1999)(Rawlings 2000).

(3) Practically, many actual processes are nonlinear but can be regarded as linear in a small operating range. Most applications can adopt linear MPC to solve complex problems. The receding horizon manner of MPC simplifies a complex system by segmenting a long-term complex process optimisation into short-term linear optimisation problems which are easier to solve (Gros et al. 2020).

(4) Besides the control theory in MPC, it is also superior to handling multiple constraints. Compared to similar control strategies, for example, PID, MPC shows a great advantage in coping with multiple constraints simultaneously on the subjects and offers better performances compared to similar control laws. MPC allows the constraints to be added both to the manipulated variables (MV, input) and controlled variables (CV, output) as well as the state (Houwing *et al.* 2008). Comparingly, PID has to run several separate control systems to add more constraints. The work of Mayne (2000) shows that MPC can systematically handle physical constraints, and its stability is thoroughly investigated. M. Pisaturo (2015) has developed a model for controlling the start-up of a passenger car using a dry clutch in an automatic manual transmission. Similarly, a vehicle control model by MPC has been conducted by Jung (2012) with constraints on the state and the input.

(5) Linearity may make system forecasting errors, therefore affecting the optimisation. However, MPC has another advantage compared to normal model-based optimisation, which is brought by its receding horizon strategy (Agrawal 2020). Under this strategy, MPC is adept at forecasting events or conditions that are closer in time or proximity to the current moment. MPC is considered to be more optimal since the proximal prediction has a smaller error compared to a distance prediction, because the prediction error is generally increased when the prediction horizon is long. Also, its feedback mechanism compensates for the estimation errors that are produced by the uncertain structure between the model estimation and the actual process. This conservatively guarantees the control is optimal. The combination of prediction and optimisation is the main difference and advantage compared to conventional control laws (Mayne *et al.* 2000b).

(6) Receding horizon strategy also provides advantages for MPC to manage the system that exists the time delay and inverse response by setting up an appropriate prediction horizon and control horizon. Time delays refer to the lag between an action taken in the system and the resulting effect, while inverse responses occur when the system reacts in the opposite direction to the applied input before eventually stabilising. By taking into account the anticipated future behaviour of the system within the prediction horizon, MPC controller facilitates the dynamic adjustment of both the prediction horizon and the control horizon. This adaptability ensures that the controller is capable of meeting the overall control objective effectively. when compared to

other feedback controllers, as it enables MPC to effectively and naturally address complex system dynamics and achieve better control performance in the presence of time delays and inverse responses (Holiš & Bobál 2015).

(7) Because of the finite control horizon and a fixed moving window, linear control problems under MPC control are easy to be transferred to quadratic programming problems, which is more computationally efficient (Rao *et al.* 2001).

However, while MPC offers indisputable advantages, some academics argued that the additional computing load brought by the design cannot be overlooked even though the system performance is rigorously guaranteed by theory (XI *et al.* 2013). This is also the main reason that the application field questioned the feasibility of MPC theories. In response to this issue, the technique of "offline design, online synthesis" is recommended in the qualitative synthesis of predictive control. By converting part of the online calculation of the integrated control law into the offline calculation, the purpose of reducing the amount of online calculation is achieved. Hu proposed an offline MPC output feedback scheme for the system with both parametric uncertainly and bounded disturbances, estimated error rates of the previous states were used to refresh that of the current online, and the result showed the joint implementation of offline and online largely saved the computational time and increased the prediction accuracy of the system (Hu & Ding 2018). Also, a light, efficient system identification model is needed to save computational costs.

Table 2-4 compared MPC to the most similar control strategy in the literature, namely Proportional Integral Derivative (PID), and provides the reason we select MPC to be the proposed optimisation scheme.

Table 2-4 The comparison of PID and MPC

Model	**The input-output variables**	**Prediction**	**Constraints**	**Specific to this study**
PID	Difficult to handle multi-input and multi-output	Control depends on past and current states but does not consider the future state.	Have to run several separate PID models to add the constraints	DC is a complex environment, multi-variables have to be estimated, and the constraints can be many. Also, a time delay is another property in the DC for one variable to respond to another. Therefore, MPC is preferred.
MPC	Easy to handle multi-input and multi-output	Control can consider the prediction of future movements and implement past decision variables to control the future state.	Can easily handle multiple constraints	

2.3 State of the art

Overall, from the literature, we can conclude that MPC is a proper scheme to control the complex environment, such as DCs, under which the optimisation can be run dynamically with optimal control strategies provided for each control horizon. To reduce the modelling effort and computational effort, a statistical linear model VAR will be applied to identify the environment. The non-convex nature of optimisation problems involving time delay and multiple variables makes them difficult to solve using traditional mathematical approaches, and the problems are known to be NP-hard. Therefore, a heuristic optimisation solver- PSO optimisation is an appropriate solver and will be applied to search for the control solutions in the feasible field. Combining these three approaches is the result of considering the advantages of each field.

The contributions of the proposed study to the literature include the following:

(1) Filled the gap in the literature by providing an energy-efficient operation control strategy of DC without changing IT and cooling devices. Offered a dynamic energy-saving managerial plan from a new perspective – the utilisation of the AC setpoints combinations.

(2) An optimal control strategy was proposed in this work, distinguishing it from other literature that mainly focused on post-event static optimisation.

(3) The proposed real-time statistical model VAR can identify the large complex system and largely save the traditional mathematical modelling effort, without a necessary deep understanding of the thermal nature, which has been wildly studied in the literature. The VAR model will be fast reacted to environmental changes, also reducing the computational power and time-consuming that has been considered an impediment in most control laws.

(4) The feasibility of MPC optimal control strategies will be validated in the actual DC environment, which can be referenced by future studies.

(5) It will be a novel attempt to combine the statistical model with the optimisation strategy and the system control strategy. Instead of the traditional mathematical approach which requires deep knowledge of the system and considerable human effort, the combination of the statistical model will bring the benefits of fast speed, efficiency, and easy implantation, moreover, easy to adapt to different complex environments.

CHAPTER 3

RESEARCH PHILOSOPHY AND PROBLEM STATEMENT

This chapter discusses the common procedures of design science studies and how the proposed study that utilises the VAR-PSO-MPC approach emphasises the design science field and follows its' procedures to solve the addressed problem.

3.1Overview of design science research philosophies

Kumar, R. (2021) defines research philosophy as a set of assumptions or beliefs that relate to the nature of knowledge and the ways it can be acquired. This essential element underpins any research project, guiding the choice of research methodology and assisting the researcher in effectively tackling the research questions.

Positivism, interpretivism, and critical theory are the three frequently employed research philosophies. As per Norris, G. (2021), positivism is grounded on the belief that there exists an objective reality that scientific methods can observe and measure. On the contrary, interpretivism asserts that reality is subjective and is not directly observable, but instead, it is constructed through social interactions and interpretation. While critical theory seeks to examine the influence of power dynamics in society on the development of people’s perceptions and comprehension of reality.

While it is possible to argue that design science (DS) research can draw upon elements of positivism, interpretivism, and critical theory, it is not typically classified within any one of these traditional research paradigms. Rather, design science research is often viewed as a distinct approach that combines elements of theory and practice in the creation of innovative artefacts or systems to address specific problems in the field of information systems (IS)(Hevner and Chatterjee,2010). Van (2004) posits that the integration of design science research methodology into the field of Information Systems (IS) complements traditional research philosophies, in essence, design science research methodology serves as a supplementary approach that can enhance conventional research paradigms in the IS field. The design science research methodology integrates theoretical concepts with practical applications to create innovative solutions that can be systematically tested and refined through iterative processes. As such, this methodology holds significant value in advancing the field of IS in operations management (OM) and beyond. Van (2013, 2016) emphasises the importance of design science research in creating novel artefacts or systems to address specific problems in the field of IS.

3.2 A design science methodology for minimising energy consumption in Data Centres

This study follows the five-stages Design Science Research (DSR) process that defined by Vaishnavi and Kuechler (2015). Within the context of this study, the adoption of the DSR approach proves essential. It facilitated a systematic framework for defining the problem domains and inspired a comprehensive progression of research steps concerning various facets of evaluation.

The flowchart shows in Figure 3-1 presents the standard procedures that govern design science studies and explains how the proposed VAR-PSO-MPC approach, adheres to these protocols.

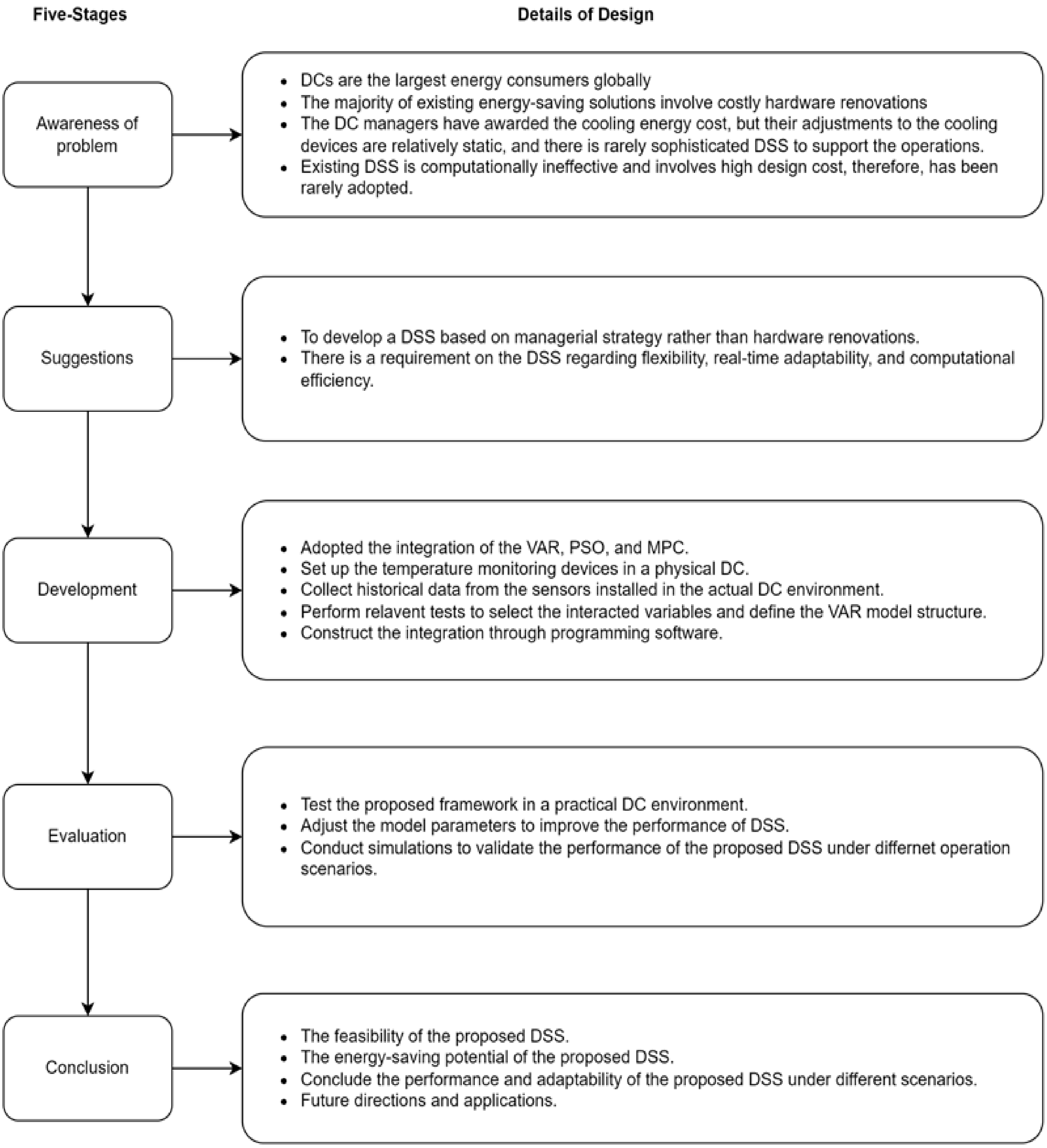


Figure 3-1The DSR process flowchart

The primary step is to identify a problem that requires a solution. Accordingly, the primary objective of the proposed study is to address the issue of high energy consumption in DCs. In the digital era, DCs play a crucial role, and their energy usage is significant, amounting to approximately 1% to 3% of the world's total energy consumption (Liu, Yu, & Wang, 2021). Additionally, existing method mainly focusing on costly hardware renovation. Thus, this study aims to develop a comprehensive methodology that can effectively minimise the energy consumption of DCs without changing existing infrastructure.

The second stage of conducting DSR involves specifying the suggested objectives of the research, which derived from the identified problem. In response to the pressing energy challenges within DCs and recognising gaps in current energy-saving approaches, our study introduces a problem-oriented solution with a comprehensive framework. A Decision Making System(DSS) should be developed based on managerial strategy rather than hardware renovation.

The third step is to design and develop the artefact or system to solve the problem. In this study, the proposed artefact is the VAR-PSO-MPC methodology. Which combines the VAR time-series modelling technique, the PSO optimisation algorithm, and the MPC control technique to create an effective solution. We designed and developed the methodology by integrating these approaches, taking into consideration their strengths and limitations to address the problem comprehensively.

The fourth step is to evaluate the artefact or system. the feasibility, scalability and robustness of implementing MPC have been assessed through a combination of practical experiment and simulation. The practical experiment involved the application of MPC control strategies within the actual DC to regulate temperature, manage energy consumption according to the variation of server's workload. Due to the mission-heavy nature of the DC, the scalability and robustness of MPC towards certain high-impact scenarios, such as additions and removals of devices, changes in server room layouts, and extreme environmental conditions could not be practically tested. To address this limitation and comprehensively evaluate the MPC system, simulation has been designed and conducted.

The final step of design science is to communicate the results of the study. The outcomes and methodology of this study have been presented in this dissertation and additional research papers are presently being pursued. Furthermore. The methodology that was developed in this study has been shared with the DC industry to aid them in reducing their energy consumption and promoting the sustainable development of society.

Overall, the proposed study that uses the VAR-PSO-MPC approach adheres to the established procedures of design science research. The VAR-PSO-MPC methodology exemplifies the design science research philosophy in several ways. It takes a problem-focused approach to design a solution for the DCs' energy efficiency problem. It combines elements of theory and practice by incorporating three different techniques to optimise different aspects of the data centre's energy usage. It creates an innovative artefact or system that addresses the problem comprehensively and effectively. Also, the findings of this study and the proposed methodology can be referenced by energy-heavy industries, making it a valuable contribution to the field of information systems.

CHAPTER 4

METHODOLOGY

This chapter gives a brief introduction to the proposed methodology and introduces the DC variables. Subsequently, the detailed formulation of the optimisation processes under MPC control to minimise the power consumption in the DC environment is presented. The chapter concludes with a process graphic containing the formulations, demonstrating the iteration of the MPC control process.

4.1Method outline

An overview process of the generalised control scheme is shown in Figure 4-1. The combination of VAR-PSO-MPC will provide a systematic dynamic optimisation process strategically. Before introducing each approach in detail, in this chapter, a basic concept of the MPC control strategy will be explained and an overview of the theoretical framework of VAR-PSO under the MPC scheme will be outlined.

There are several components have been involved in the entire process, including the following:

(1) A system identification model.

(2) A prediction mechanic for the future system.

(3) An optimisation solver to optimise the cost function for each forecasting horizon.

(4) The plant, to which the optimal control strategies will be implemented.

MPC is an optimisation method to solve the control problem. In the loop of the MPC controller at the bottom of the graph, a fixed system model will be constructed to predict the future movement of the predefined certain time horizon. Then the optimiser will optimise the objective function (would be the cost function of MPC) to solve the optimisation problem (minimise the gap between the manipulated system status and the referenced system status) time by time on each time step over the prediction horizon, this optimisation process will be done under a receding horizon manner. As for the implementation, although the optimiser will optimise the objective functions and provide a sequence of optimal control actions correspondingly, in practice, the manager will only take the first action in the optimal sequence to apply to the plant (This can be seen from the upper loop in Figure 4-1). The performance of the plant will be recorded as a new system condition and will be applied for the new iterations for the next prediction and optimisation. The output from the optimisation act as the input to the plant and changes the system status practically, therefore, also act as an input to the prediction and optimisation for the next loop. The strategy guaranteed the optimal control conservatively because this optimises the system with the consideration of its future performance while preventing unpredictable changes in the future horizon.

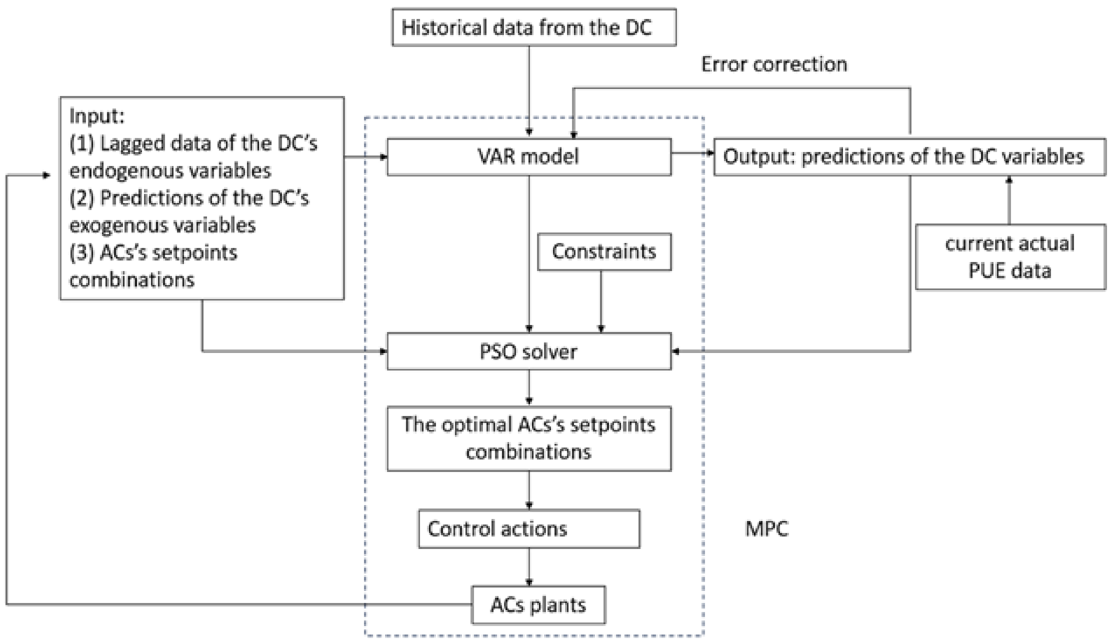


Figure 4-1 The processing diagram of VAR-PSO-MPC framework

The combination of VAR-PSO-MPC provides a systematic dynamic optimisation process strategically. The VAR model, as a multivariate time series model, captures the intricate relationships among the system variables, incorporating lagged values and multi-directional interactions. It takes the real-time data as input to forecast the future movements of the interacted variables in the DC room. Based on VAR prediction structure, the PSO algorithm solves the objective function by manipulating the ACs temperatures setpoints and minimise the gap of forecasted PUE value and desired PUE value. Within the MPC control framework, the optimised control actions are implemented at appropriate intervals, and real-time feedback from the system will be utilised as an input to VAR prediction of the next iteration. Furthermore, error correction is performed according to the system feedback, and VAR parameters are adjusted at each time step.

This strategic integration addresses challenges and fills existing gaps through the following key aspects:

(1) Inherent Complexity Navigation: The VAR model excels in capturing intricate relationships within complex environments. And it also superior in demonstrates computational efficiency.

(2) Complexity Reduction Efforts: We employed Granger causality test, Impulse response function, and Variance Decomposition within the VAR modelling framework. These methods assist in variable selection and offer a comprehensive understanding of the complex interactions within the DC. By identifying key variables and their interdependencies, the dimension of the optimisation problem has been reduced. This reduction makes the optimisation space more manageable. Unlike conventional optimisation systems that

directly integrate complexities into the objective function, our DSS strategically incorporates this complex relationship during the VAR estimation stage, effectively simplified the optimisation algorithm. This reduces computational loads significantly. Furthermore, MPC frameworks segment the complex nonlinear problem into smaller linear problems at each time-step, further reduced the complexity and enhanced the system efficiency.

(3) Temporal Dynamics Consideration: The VAR model accounts for delays in variable responses, crucial in the context of DCs. Rigorous criteria such as the Akaike Information Criterion (AIC) and Schwarz Criterion (SC) are employed to determine the appropriate lag length, enhancing the accuracy of capturing temporal dynamics.

(4) Real-Time Adaptability: To accommodate the dynamic real-time environment, our approach integrates synchronised time series data into the VAR model, facilitates real-time feedback from the system and enabled error correction.

(5) Increased optimal performance: The incorporation of PSO algorithm improved the efficiency in searching the global optimal solutions and significantly reduced the possibility to fall into local optima.

It is Noteworthy that the integration of three approaches will not only leverages their respective strengths, but also compensating for their individual limitations and maximising the synergistic potential. This strategic integration also highlights our methodological innovation, differentiating our approach significantly from conventional practices in comparable industries. Table 4-1 shows their individual strengths and limitations, as well as how the limitations have been compensated by integration.

Table 4-1 Individual Strengths, Limitations, and Compensations through Integration"

	Strength	Limitations	Benefit from integration
VAR	• Captures interdependencies among variables • Enhances interpretability of complex systems • Computational efficiency	• Long-term prediction accuracy	• MPC allows for real-time refinement using sensor data • MPC enables complex problem to be segmented into small linear problems, therefore reduce error
PSO	• Superior at exploring large solution space • Finding global optima • Simplicity and flexibility	• Influenced by initial particle positions • Struggles in high-dimensional problems • Temporal dynamic of the system may increase complexity for PSO to handle	• VAR simplifying this complexity • Refined initial starting point based on VAR prediction and MPC real-time feedback • VAR prediction offering PSO insights into the future enables PSO to handle the temporal dynamic naturally
MPC	• Ability in handling multiple objectives and constraints • Considered future performance of the system • Real-time adaptable	• Conventional MPC relies solely on predefined system models and gradient-based optimisation methods, required significant design efforts	• VAR as a data-driven approach, provides the system model without the need to study the complex system physics. • PSO explores the solution space more comprehensively than traditional gradient-based methods

4.2 MPC model form

4.2.1 The Basic Assumptions for MPC

Because the MPC approach emerged from industry practices, therefore there are assumptions not only about the models but also about the plants. The model assumes the environment is stable and follows a certain manner. Therefore, the data collection was assumed to be no failure, and the devices work in normal mode.

Assumption I: No sensor failure or noise.

Assumption II**:** The system identification model is assumed to be appropriate and estimation errors are negligible.

Assumption III**:** No unpredicted environmental disturbances.

Theoretical formulation of MPC
Minimise the cost function, which is the summation of the state cost function: $\underset{u(k),\ldots u(k+N-1)}{Min.} J_N(X_N, U_N) = \sum_{i=1}^{N}(x(u)_{k+i} - x_{ref})^2$ Subject to   Model of the process innovation $x(k+1) = f(x_k, u_k)$ States' constraints $x_k \in X_k, k = 1, \ldots, N-1$ Actions' constraints $u_k \in U_k, k = 1, \ldots, N$

The objective function is the summation of the stage cost functions on each time step over the prediction horizon. The stage cost functions are the deviations between the model's predicted future states and the reference states. The objective function will be subject to the system identification model, which is the innovation of the state and control actions. Also, there would be other constraints on both control actions and the states, according to the industry conditions.

Figure 4-2 shows the control principles of MPC. Three basic components of the MPC (prediction, online optimisation, and receding horizon operation) are demonstrated. The lower part of the graph demonstrates the changes in the control actions and the upper part of the graph demonstrates the changes in the system state in response to the changes in control actions. By predicting the future objective functions, the optimal control actions will be given by the optimisation algorithms. The optimal control actions will be only implemented for a short period, and the system will respond to the control actions applied

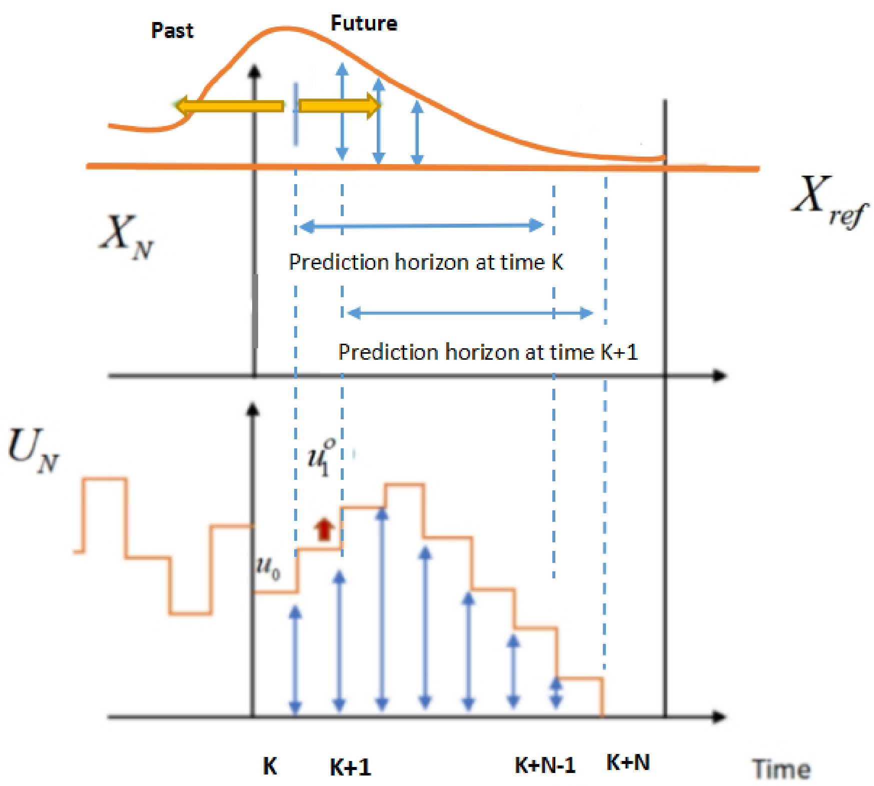


Figure 4-2 MPC control principle

to the devices. Optimised control actions will be applied recursively. With the optimisation window marching on, the full control horizon can be optimised.

4.3 Optimisation problem formulation

- **The control actions u_t**

Let AC temperature setpoints combination at time k denoted as vector u_k, according to the ACs' attributive:

$$u_t = T_{v,t}^{cool} = \begin{bmatrix} T_{1,t}^{cool} \\ T_{2,t}^{cool} \\ \dots \\ T_{v,t}^{cool} \end{bmatrix}^T, \forall t \in T, T_{v,t}^{cool} \in [16, 28] \tag{4.1}$$

Where the range was given by the ACs' setting temperature points in the plants' specifications.

- **The state x_t**

The PUE index as one of the endogenous variables in VAR model, which belongs to the endogenous set $Y_t = (y_{1t}, \dots y_{mt}), M \in \mathbb{N}$, also serves as the system state in MPC, therefore can be written as $PUE = y_{1t} = x_t$. The PUE value closer to 1, the more efficient the DC operates.

According to the concept of MPC, the **stage cost function** will be denoted as:

$$\ell(x, u) = \left\| x(u)_t - x_{ref} \right\|^2 \tag{4.2}$$

This represents the stage gap between the system state after operating the control action and the expectation of the state (the reference state).

The **objective function (Cost function) of MPC** would be the summation of the stage cost functions over the whole prediction horizon from time $t = k \ to \ k + N$:

$$\underset{u(k),\dots u(k+N-1)}{Min.} J_N(X_N, U_N) = \sum_{i=1}^{N} (x(u)_{k+i} - x_{ref})^2 \tag{4.3}$$

Subject to the constraints:

$x_{k+1} = f(x_k, u_k)$, where $f(x)$ is the system prediction function, which will be formulated by the VAR model in the next section.

$x(0) = x_0$, where$x_0$is the initial value of the state. Can be obtained through historical data.

(1) $u_k \in U, \forall k \in [0, N-1]$,
(2) $x_k \in X, \forall k \in [0, N]$.

The **value function** is the minimum value of the objective function (cost function):

$$V_N(x) = \underset{u_k,\dots u_{(k+N-1)}}{Min.} J_N(X_N, U_N) = \sum_{i=1}^{N} (x(u)_{k+i} - x_{ref})^2 \tag{4.4}$$

4.4 Solution structure with VAR model

4.4.1 Variables specifications and model assumptions

VAR is a data-driven model for system identification. As a time series model, it also is used as the prediction function in the MPC control process. In this section, the optimisation theory of MPC control under the solution structure with the VAR model will be demonstrated.

Table 4-2 lists the DC variables that involve in the server room environment. Variables are categorised into two groups: endogenous variables and exogenous variables. These variables are influenced by changes or fluctuations within the environment itself, at the same time, their behaviours also affect to the environment. ACs' outflow temperature measured from the closed point of the airway ($T_t^{outflow}$:), server room ambient temperature ($T_t^{ambient}$), the heat generated from the servers that measured by the temperature difference of the front and back side of the server rack (HG_{jt}), and PUE value are considered to be endogenous variables. Exogenous variables, on the other hand, are external factors that have an impact on the server room environment but are not influenced by changes within the environment itself. The temperature outside the DC ($T_t^{outdoor}$), the server's given IT workload (WL_t), and the ACs temperatures setpoints temperatures ($T_{v,t}^{cool}$) are exogenous variables, which affect the system but are not affected by the system.

Table 4-2 The DC variables list

Endogenous variables	Exogenous variables

$T_t^{outflow}$	$T_t^{outdoor}$
$T_t^{ambient}$	$T_{v,t}^{cooling}$
HG_{jt}	WL_t
PUE_t	

- **Assumptions**

The five basic assumptions for the VAR-based MPC control are shown as the following:

Assumption I: The system is linear in each time step.

Assumption II: Each control action $u_{(t)}$ remain constant in each implementation horizon.

Assumption III: DC environment is a complex environment with DC components mutually affecting each other. In this study, all the indoor temperature variables and energy consumption variables are modelled as endogenous variables, and the given IT load, AC setpoints temperatures, and outside temperature are modelled as exogenous variables.

4.4.2 VAR model form

General VAR(p) model consist of *M* endogenous variables Y_t, and $p = P_{max}$ time-lags, with exogenous variables Z_t denoted as:

$$VAR(p): Y_t = A_1 Y_{1,t-1} + \cdots + A_p Y_{M,t-P_{max}} + BZ_t + \varepsilon_t \tag{4.1}$$

The length of the lag will be selected by AIC and SC. Due to time-series model's characteristic, VAR model would predict the states and actions for future *t* + *δ* time periods ahead recursively. The following sessions will demonstrate how VAR model been utilised to predict the future states.

4.4.3 The VAR model prediction

In this section, the prediction method based on the VAR model will be demonstrated.

Take VAR(1) as an example to be simplified:

$$VAR(1): Y_t = \Phi_1 Y_{t-1} + BZ_t + \varepsilon_t \tag{4.2}$$

Z_t is a set of exogenous variables (including the control actions $u_{v,t} = T_{v,t}^{cool}$, the DC outdoor temperature $T_t^{outdoor}$, and the given IT workload WL_t), where the AC temperature setpoints are given by the optimiser, and two separate ARIMA forecasting models will be applied to forecast the outdoor temperature and the CPU usage for the target forecasting period. Therefore, for the VAR model, $\mathrm{Z_t}$ can be regarded as the deterministic regressors that are given. Let $\mathrm{BZ_t} = \Psi_1 \mathrm{u_t} + \mathrm{c}$, represent the exogenous term $\mathrm{BZ_t}$ is a function of the control actions and a constant. Assume that the ACs temperature setpoints vector is adjusted at time t and stays constant in each forecasting horizon until time $t+\delta$. The innovation of the forecasting is shown as function (4.3) as follows:

$$\begin{aligned} Y_t &= A_1 Y_{t-1} + B_1 u_t + c + \varepsilon_t \\ Y_{t+\delta} &= A_1 Y_{t+\delta-1} + B_1 u_t + c + \varepsilon_{t+\delta} \\ Y_{t+\delta-1} &= A_1 Y_{t+\delta-2} + B_1 u_t + c + \varepsilon_{t+\delta-1} \\ Y_{t+\delta} &= A_1(A_1 Y_{t+\delta-2} + B_1 u_t + c + \varepsilon_{t+\delta-1}) + B u_t + c + \varepsilon_{t+\delta} \\ &= A_1^2 Y_{t+\delta-2} + (A_1 + I) B_1 u_t + (A_1 + I)c + B_1 \varepsilon_{t+\delta-1} + \varepsilon_{t+\delta} \\ &= \cdots \\ &= A_1^{\delta} Y_t + \left(\Phi_1^{\delta-1} + I\right) B_1 u_t + \left(A_1^{\delta-1} + \cdots + A_1 + I\right)c + A_1^{\delta-1}\varepsilon_{t+1} + \cdots \\ &\quad + A_1 \varepsilon_{t+\delta-1} + \varepsilon_{t+\delta} \end{aligned} \tag{4.3}$$

Then the conditional expectation (mean) will be as the following equation:

$$E(Y_{t+\delta}) = A_1^{\delta} Y_t + \left(A_1^{\delta-1} + I\right) B_1 u_t + \left(A_1^{\delta-1} + \cdots + A_1 + I\right)c \tag{4.4}$$

Substitute the DC variables into Extend the above equation to be extended to a matrix form:

$$E\begin{bmatrix} PUE \\ HG \\ T^{outflow} \\ T^{ambient} \end{bmatrix}_{t+\delta} = A_1^{\delta} \begin{bmatrix} PUE \\ HG \\ T^{outflow} \\ T^{ambient} \end{bmatrix}_{t} + (A_1^{\delta-1} + I)B_1 \begin{bmatrix} T_1^{cooling} \\ T_2^{cooling} \\ \ldots \\ T_v^{cooling} \end{bmatrix}_{t} + (A_1^{\delta-1} + \ldots + A_1 + I)c \tag{4.5}$$

Where the constant term c is practically calculated by the predictions of exogenous variables $T^{outdoor}$ and workload WL_t over the prediction horizon $t_0 \leq$

$t \leq t_{predict}, t \in T$, are predicted by the ARIMA model.

Extract the function of the state variable $x_t = PUE$ as and decision variable and $u_t = [T_{1,t}^{cooling}, T_{v,t}^{cooling}]$, from equation (4.5), then the system innovation function can be represented as the following equation:

$$x_u(t): E\,(x_{t+\delta} = a_1 x_t + a_2 HG_t + a_3 T_t^{ambient} + a_3 T_t^{outflow} + B_1 u_t + CPU_t + T_t^{outdoor} + c \quad (4.5)$$

This can be easily expanded to $VAR(p)$ case. And the order of the VAR model p can be selected by AIC and SC criteria.

The optimisation problem then can be formulated following the MPC formulation law.

Let k denote the optimisation start point of each time horizon, for each horizon $t = k + N$, the objective function is to minimise the total running cost of each stage:

$$\underset{u_1 \ldots u_k}{Min.} J_N(X_N, U_N) = \sum_{i=1}^{N} [x(u)_{k+i} - x_{ref}]^2 \quad (4.6)$$

Subject to the following constraints:

s.t.

a) $x_{k+1} = f(x_k, u_k)$ (4.7)

where $f(x)$ is the system prediction function formulated by the VAR model.

b) $\forall t = k, 18 \le T_t^{ambient} \le 27$ (4.8)

This is the restriction of room temperature provided by ASHRAE Guideline 2016

c) $\forall t = k, 16 \le T_{v,t}^{cooling} \le 28$ (4.15)

The restriction of temperature setpoints from the AC attributive

d) $\forall t = k, 0 \le WL \le 1$

A workload is a percentage number between 0 and 1 which is the CPU usage of the servers. (4.16)

e) $\forall t = k, 1.2 \le x_t \le 2.5$

The constraint on the state variable. According to the practical environment and hardware condition in the target DC. which is the PUE value should be a number between 1.2 and 2.5. (4.17)

The equation of the above constraints can be derived from $VAR(p)$ process equation.

4.4.4 Error evaluation and error correction

For the general recursive prediction, only one model with fixed coefficients will be applied to the whole forecasting horizon, and the real past information retrieved from the environment to input to the forecasting model would be only used to forecast the first period. From the second period, all the past information input to forecast is based on the previous forecasted value rather than the real information from the environment. This will accumulate the forecasting error exponentially. The direct prediction method can reduce the forecasting error by re-estimate the model coefficient before each prediction starts. However, this would be computationally costly. The proposed study adopted the error evaluation and error correction method for each prediction horizon.

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y' - y)^2} \tag{4.18}$$

Root mean square error (RMSE) accuracy will be applied to evaluate the prediction accuracy of time $t - 1$ at time t.

Eq.19 is the formulation of RMSE, where y' is the forecasted value of the variable and y is the measured value of the variable at time t.

The system evaluates the prediction error for the previous stage at every start point of the new prediction. The error bound is set to 0.4. if RMSE is higher than 0.4, the system will retrieve the latest historical data from the data API and re-estimate the VAR model coefficient.

4.5 The optimal control for the DC environment

Suppose that we have n sets of AC outflow temperature combination vector

u_k, denotes ACs' setpoints combinations of all AC units at time k. Then the searching space of the AC temperature set point combinations will be denoted as follows:

$$Z^n = \begin{bmatrix} u_{1,k} \\ u_{2,k} \\ \cdots \\ u_{n,k} \end{bmatrix}, n \in N^*. \qquad (4.19)$$

Let $\rho: Z \to L_x(Z^n)$, maps the control actions sets of temperature combination bundles into a permutation matrix Z. Where $Z^n$is the arbitrary field.

As a result, a traditional optimisation approach that substitutes all the feasible solutions of size η into the optimisation equation will be impossible. Because in that case, a great computational effort will be required to calculate the $F(u) = \left(f_1(u), f_1(u), \dots . f_\eta(u)\right)$ at every time step to obtain the optimal solution u_k^o.

To solve the optimisation problem PSO approach will be adopted. The PSO algorithm is an evolutionary algorithm. It states with a random solution and finds the optimal solution by iteration. And then testing the solution quality by fitness value.

Suppose we have, each particle i has the velocity $V_i$and position χ_i, χ_i is a γ dimensional vector that represents the AC temperature setting range

$[\tau_{low}, \tau_{high}]$.

In PSO optimisation theory, we will have:

$$v_i = \varpi v_i + c_1 \times rand() \times (pbest_i - x_i) + c_2 \times rand() \times (gbest_i - x_i) \tag{4.20}$$

$$x_i = x_i + v_i \tag{4.21}$$

In the above equations,

$i = 1,2, \ldots, n$, where n denotes the number of particles. Here would be the total number of ACs temperature setpoints comb;

v_i is the velocity of the particle i.Here would be the AC temperature setpoint changes and the direction (increase/decrease) of each change.

V_{max} the maximum value of v_i $v_i < V_{max}$,if $v_i > V_{max}$, then $v_i = V_{max}$.

$rand()$ is a random number between (0,1);

x_i is the current position of the particle i.

c_1, c_2 are learning coefficients, normally set to $c_1 = c_2 = 2$

ϖ is the momentum coefficient. A greater ϖindicates a higher ability in seeking global optimum but a lower ability in seeking partial optimum. Vice versa.

The optimisation process will be demonstrated as follows:

Step1:

Initialise:

$$P_1 = \begin{cases} V_1 = (\Delta T_{1,1}, \Delta T_{1,2} \ldots \Delta T_{1,l}) \\ X_1 = (T_{1,1}^{\sup}, T_{1,2}^{\sup} \ldots T_{1,l}^{\sup}) \end{cases} \rightarrow \begin{cases} f_1 = f(X_1) \\ pBest_1 = X_1 \end{cases} \tag{4.9}$$

$$P_2 = \begin{cases} V_2 = (\Delta T_{2,1}, \Delta T_{2,2} \ldots \Delta T_{2,l}) \\ X_2 = (T_{2,1}^{\sup}, T_{2,2}^{\sup} \ldots T_{2,l}^{\sup}) \end{cases} \rightarrow \begin{cases} f_2 = f(X_2) \\ pBest_2 = X_2 \end{cases}$$

$$P_n = \begin{cases} V_n = (\Delta T_{n,1}, \Delta T_{n,2} \ldots \Delta T_{n,l}) \\ X_n = (T_{n,1}^{\sup}, T_{n,2}^{\sup} \ldots T_{n,l}^{\sup}) \end{cases} \rightarrow \begin{cases} f_n = f(X_n) \\ pBest_n = X_n \end{cases}$$

$$gBest = min.p\ Best \tag{4.10}$$

Step 2:

Update the positions of the particles

$$\mathrm{P_1} = \begin{cases} \mathrm{V_1} = \omega * \mathrm{V_1} + \mathrm{c_1} * \mathrm{r_1} * (\mathrm{pBest_{min}} - \mathrm{X_1}) + \mathrm{c_2} * \mathrm{r_2} * (\mathrm{gBest} - \mathrm{X_1}) \\ \mathrm{f_1}^{\mathrm{a}} = \mathrm{F}(\mathrm{X_1}), \qquad \mathrm{X_1}^{\mathrm{a}} = \mathrm{X_1} + \mathrm{V_1} \end{cases}$$

$$\mathrm{P_2} = \begin{cases} \mathrm{V_2} = \omega * \mathrm{V_2} + \mathrm{c_1} * \mathrm{r_1} * (\mathrm{pBest_{min}} - \mathrm{X_1}) + \mathrm{c_2} * \mathrm{r_2} * (\mathrm{gBest} - \mathrm{X_2}) \\ \mathrm{f_2}^{\mathrm{a}} = \mathrm{F}(\mathrm{X_2}), \qquad \mathrm{X_2}^{\mathrm{a}} = \mathrm{X_2} + \mathrm{V_2} \end{cases}$$

$$\ldots \tag{4.11}$$

$$\mathrm{P_n} = \begin{cases} \mathrm{V_n} = \omega * \mathrm{V1} + \mathrm{c_1} * \mathrm{r_1} * (\mathrm{pBest_{min}} - \mathrm{X_1}) + \mathrm{c_n} * \mathrm{r_n} * (\mathrm{gBest} - \mathrm{X_n}) \\ \mathrm{f_n}^{\mathrm{a}} = \mathrm{F}(\mathrm{X_n}), \qquad \mathrm{X_n}^{\mathrm{a}} = \mathrm{X_n} + \mathrm{V_n} \end{cases}$$

Step 3:

Evaluate the particles: update $pBest$ and $gBest$

if $f_i^a < f_i$, then $pBest = X_i^a$, if not, keep $pBest_i = X_i$.

$$gBest = min. p\, Best_i \tag{4.12}$$

Step 4: Until the last iteration, end up the process, and get the most approximate optimal solution.

Procedure PSO

For each particle i

Initialise velocity Vi and position Xi for particle

Evaluate particle i and set $pBesti = Xi$

End for

$gBest = \min\{pBesti\}$

While not stop

For $i = 1$ to N ,

Update the Volocity and potition of particle i

Evaluate particle i

If $fit(Xi) < fit(pBesti)$

$pBest = Xi$

If

$fit(pBesti) < fit(gBesti)$

$pBesti = gBesti$;

End for

End while

Print gBest

End Procedure

The below pseudocode demonstrated the PSO procedure:

In the end, the system identification model, VAR prediction, and PSO optimization solver will be integrated into an MPC process. The MPC **pseudocode** can be derived as follows:

```
Algorithm :

Given:

System structure

    sys.A      :  dynamics matrix A
    sys.B      :  input matrix B
    sys.Q      :  state cost matrix Q
    sys.R      :  input cost matrix R
    sys.xmax   :  state upper limits x_{max}
    sys.xmin   :  state lower limits x_{min}
    sys.umax   :  input upper limits u_{max}
    sys.umin   :  input lower limits u_{min}
    sys.n      :  number of states
      sys.m    :  number of inputs

MPC parameters (params structure):
```

params.N : MPC horizon N

params.Qf : MPC final cost Q_f

params.kappa : Barrier parameter

params.niters : number of newton iterations

params.quiet : no output to display if true

Step 0: For the target terminal horizon $N > 0$, set the time step resolution $\delta > 0$and forecast horizon$N\delta > 0$, subdivide the time interval $[0, N]$ as

$$0 < \delta < 2\delta \ldots < (N-1)\delta < N\delta$$

Repeat

Step 1: For every forecast horizon, read the current state of the process $x(0) = x_0$,

Step 2: Compute an optimal control sequence by solving the MPC optimisation problem eq.

Solution: $u(k), u(k+1), u(k+2), \ldots u(k+N-1)$

Step 3: Apply to the plant only the first element of such a sequence: $u^o(k) = u_k$

Step 4: Update the time $k \leftarrow k+1$

Update the initial state $x_0 \leftarrow x(u^o)$

Back to Step 1

Figure 4-3 demonstrated the whole process of VAR-PSO-MPC control. The whole MPC scheme was separated into two parts: The offline process and the online process. This is for the purpose to save computational power and computational time. According to Figure 4, the VAR-PSO-based MPC control will follow the five steps below:

Step I: In the offline process, we use the historical data collected from the DC-installed sensors to estimate the VAR model parameters.

Step II: Prediction. VAR model will be used to predict future movement, starting from an initial control action u_0. The online process will start with VAR giving predictions of the future states x_ks (Future PUE value) based on the initial control actions u_0 (The initial AC's temperature setpoints) from time $0 \neg k$. The total prediction horizon is time $t = k$ to $t = k + N - 1$.

Step III: Optimisation. PSO optimisation algorithm will be applied to minimise the cost function and search for the optimal solution among ACs set point combinations under the given constraints to the state and control actions. The optimisation will be made at the time $t = k$ to minimise the summation of the future deviations between the reference value$x^r$and the predicted value in the target time horizon. This will give us the suggested optimal future control actions over the prediction horizon N. PSO optimisation solver is applied to solve the optimisation at each time step.

Step IV: MPC scheme will only take the optimal control actions at time step 1 to implement the control action at the first time step, written as u_1^o.

Step V: The first output of optimal control u_1^o from the optimisation along with the measured states will be used as an input to the next prediction, in this case, the system innovation can be achieved.

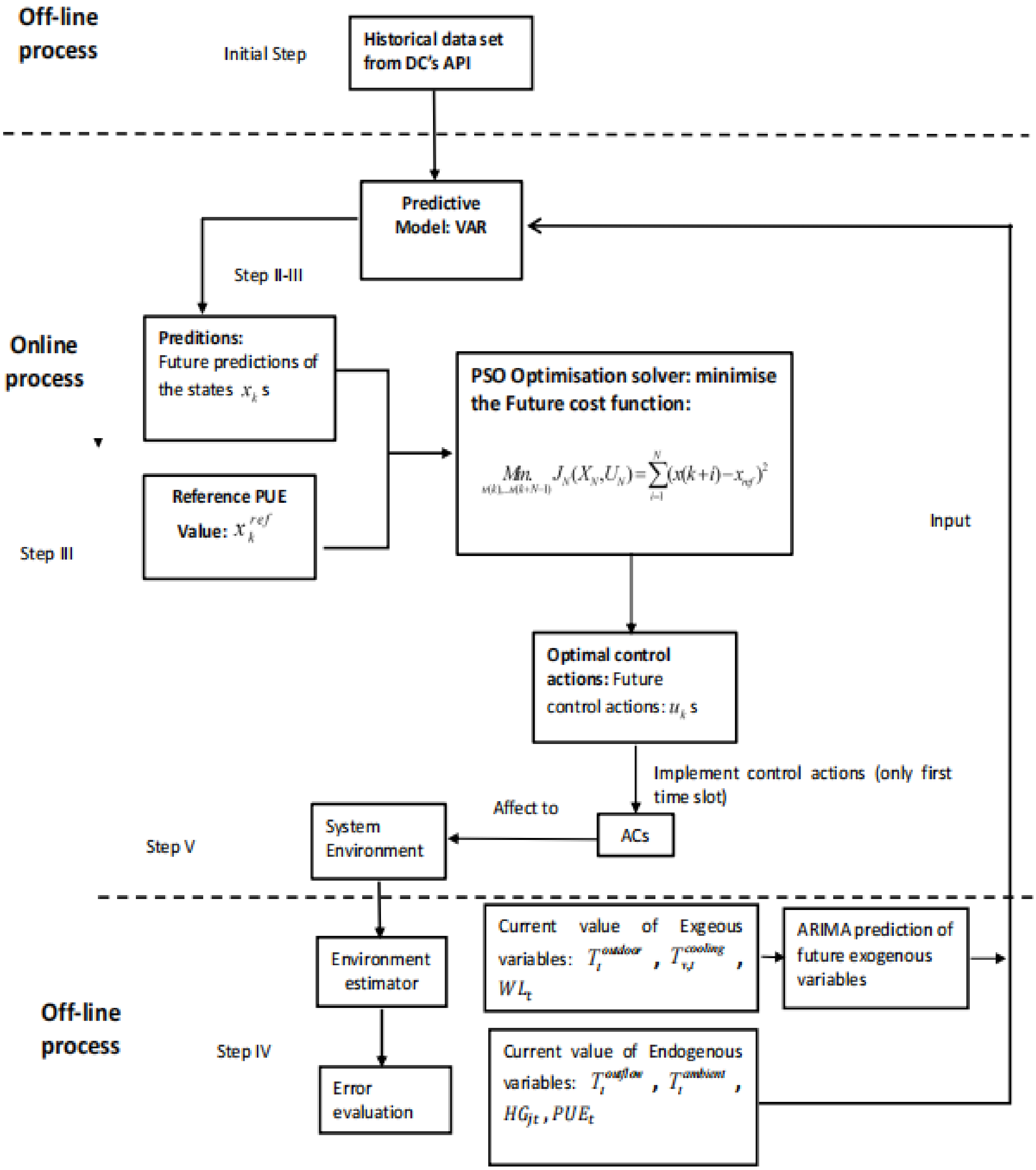


Figure 4-3 The processing graph of the VARVAR-PSO-MPC framework

CHAPTER 5

EXPERIMENT

In this chapter, the experimental results obtained will be presented and discussed. Firstly, the design of the experiment will be detailed in terms of the research objectives. Subsequently, configurations of the experiment will be presented. Following this, the data's analytical results will be thoroughly demonstrated. Based on the analytical result, the system model VAR and its' prediction performance will be evaluated. Furthermore, optimisation under MPC control will be operated and compared to several different scenarios as well as the performance of the Genetic Algorithm (GA).

5.1Design of Experiment

The experiment was carried out in a real data centre in Thailand. The experiment was divided into three parts: The first stage entails the installation of sensors, data collection, data analysis, system identification modelling, prediction validation, and model adjustment. The second stage entails testing the optimisation solver and the MPC prototype, the parameters adjustment, and performance improvements. The third stage entails performance validation and evaluation.

(1) Sensors installation

The sensors were provided by a third party, including temperature sensors on the front and back rack of the servers, sensors installed on the ceiling to detect the ambient temperature, also sensors outside the DC to detect the outside environment temperature. The third party will provide the thermal mapping service including sensors installation and management, data storage, room thermal dynamic mapping, and provide the data stream collected by the sensors. The server's workload data will be provided by the DC central management system. Also, energy consumption data is provided by the smart energy metres installed in the server room.

(2) Data collection

Historical data have to be collected for a week with one hour's resolution from 13th June 2022 to 19th June 2022. To ensure diversity of the data, AC temperature settings will be varied from 18 to 29 degrees, alternated among four AC units. During one week, there are three times of stress tests on the servers, which will manipulate half of the servers to a full workload. All the data have to be unified in a time-series form and prepared to be used to estimate the system model.

(3) Data analysis and system identification modelling

Data analysis involves estimating the stabilities of the variables, analysing the thermal environment in the server room, and detecting the causal relationship between DC components. After processing the data, the coefficients of the VAR model will be estimated.

(4) Prediction validation and model adjustment

The prediction has been conducted based on the system model VAR every 6 hours. Validation of the prediction accuracy has been conducted to tune the VAR parameters' values such as adjusting coefficients' weights or adjusting the forecast horizon.

To validate the prediction accuracy and adjust the coefficients. The data was divided into two sets: training and validation. The training dataset comprises about 75% of the data needed to estimate the VAR coefficients, while the validation dataset contains the remaining 25% to check the prediction's accuracy. Human monitoring has been seriously executed to prevent the violence of the constraints and prevent the extreme temperature settings that would lead to harm to the facilities.

(5) Testing optimisation solver and the MPC prototype

This experiment has been conducted from 20th June 2022 to 3rd July 2022.

The MPC optimal control prototype has been applied to the server room. DC manager adjusts the AC setpoints temperature according to the optimal solutions given by the optimisation solver. The implementations were following the MPC control law. Different control horizons have been applied to compare the performances.

(6) Monitoring of the temperature boundary and emergency alarm

According to the ASHRAE guideline, the recommended servers' working temperature condition should be between 15 to 32 degrees. The servers in the target server room have a down-locking system that would automatically shut down to prevent overheating and also down-lock to reduce the heat. Additionally, the DC managers have been informed to closely monitor the heat by the temperature sensors.

(7) Performances validation and result evaluation

The final validation and result evaluation have been conducted from 4th July 2022 to 10th July 2022. Validation of the control application will be compared with the result in the absence of the MPC control. Also, the result has been compared with the results from a different non-parametric approach-Genetic algorithm (GA).

5.2 Server room configuration

Figure 5-1 and Figure 5-2 show the 3D view and top view of the server room configurations. Two server rack rows are located back to back parallelly, with an underground cold air supply on both front sides of the racks. 3 AC units have been installed in a separate room, with an underground cooling aisle connected to the server room to supply the cooled air.

Table 5-1 lists the environmental and power consumption monitoring devices that have been installed in the server room. 2 temperature sensors have been installed on the ceiling to detect the room's ambient temperature. 4 sensors on the server rack front and rear side to detect the server rack temperatures. 1 sensor is attached close to the ACs outlet underground air aisle to detect the ACs outflow temperatures. 1 sensor on the outdoor chiller to detect the temperature outside the DC. Energy analysers have been installed on each AC unit and on the outdoor chiller for calculating the cooling power consumption. Smart PDUs have been installed in each target server rack. These devices are capable of measuring all sockets. To calculate the total load on server racks, the load that comes from the PDUs should be gathered.

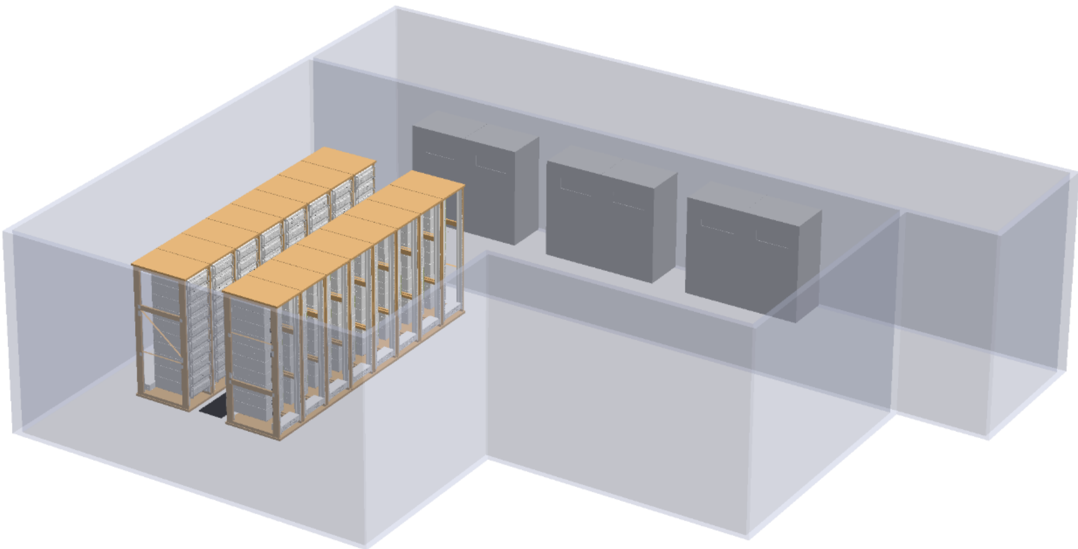


Figure 5-1 3D view of the server room

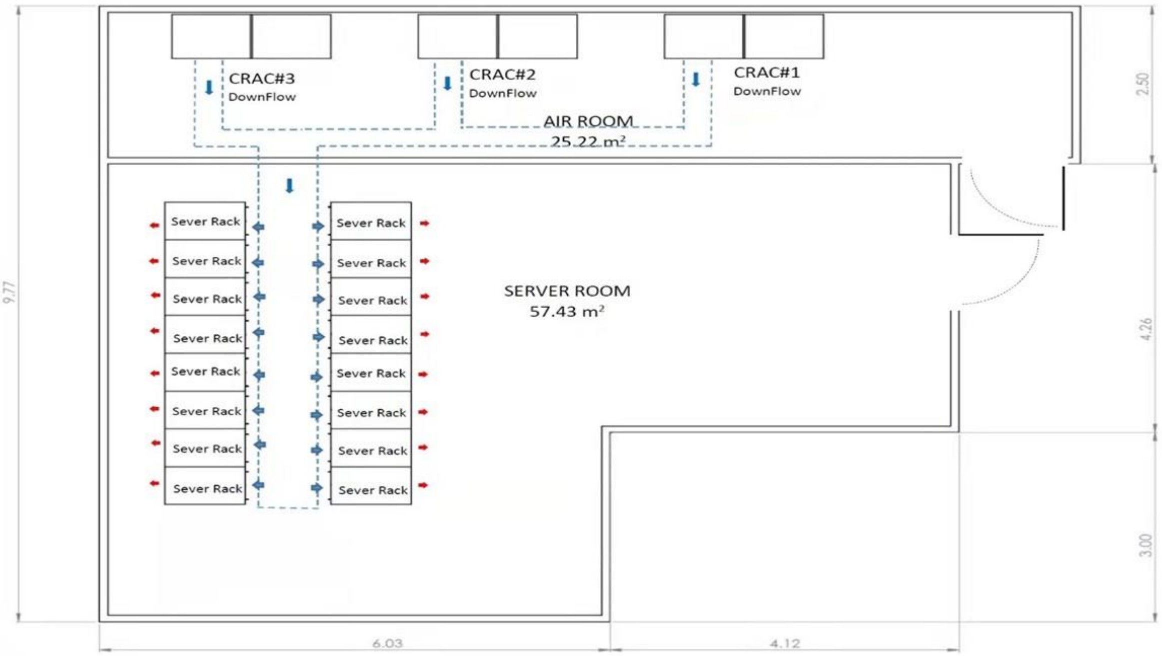


Figure 5-2 Top view of the server room

Table 5-1 The environmental and power consumption monitoring devices

Devices name	**Count**	**Location**
Temperature sensors	2	ceiling
Temperature sensors	1	Outdoor chiller
Temperature sensors	1	AC outlet air aisle
Temperature sensors	4	Server racks
Smart PDUs	14	Server racks
Energy analysers	3	AC units
Energy analysers	1	Outdoor chiller

5.3 Historical data and analytical results

5.3.1 Historical data

The historical data for one week contains 168 data points with 15-minute intervals that have been collected from an actual DC's data API in Thailand. These data have been used to model the system identification model VAR. The raw data from the sensors are in 15-minute intervals, to be adapted to the optimisation and MPC control process, it has been extracted to be one-hour intervals by sequence.Figure 5-3 to Figure 5-7 demonstrated the historical data of the DC variables according to experiment week I (Figure 5.3 shows the one server rack's front and back temperatures, and the deviation between them). Characteristics of DC variables will be analysed.

■ The historical temperatures data

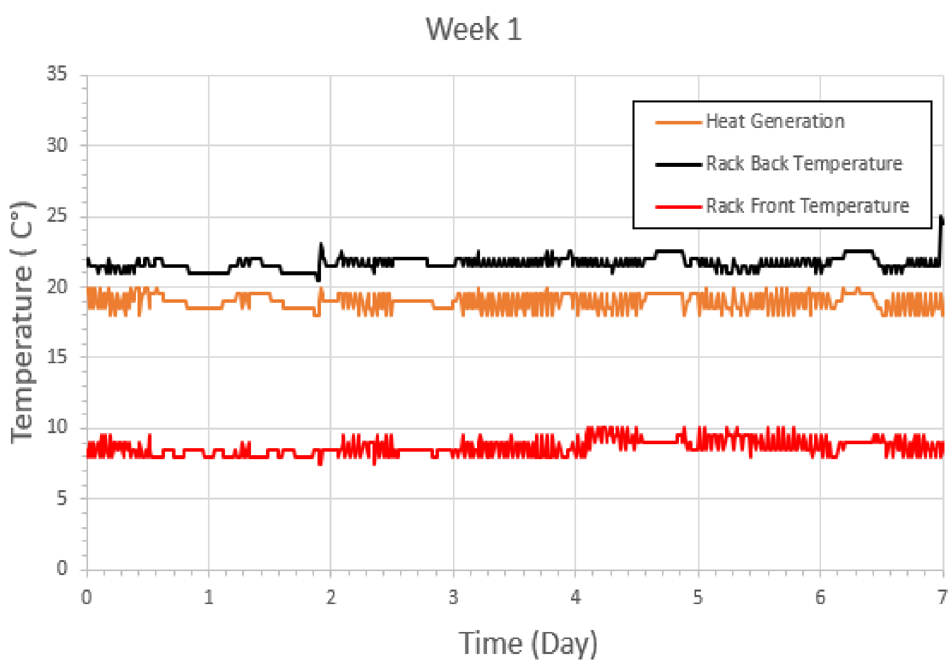


Figure 5-3 The historical data of the server racks temperatures in week 1

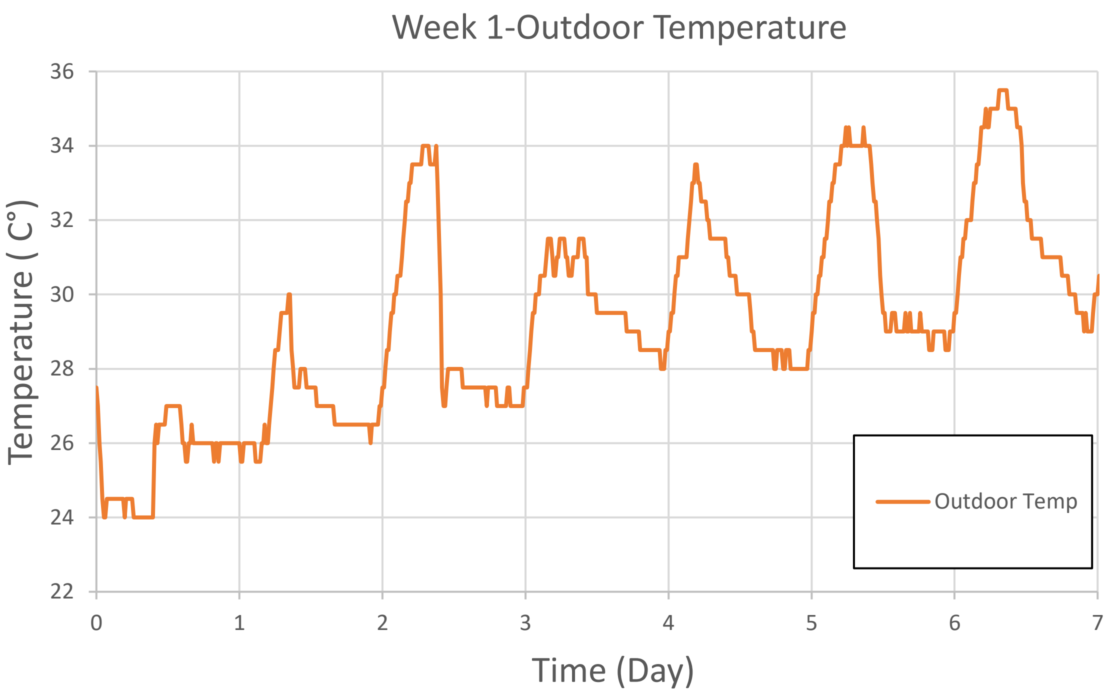


Figure 5-4 The historical data on the outdoor temperature in week 1

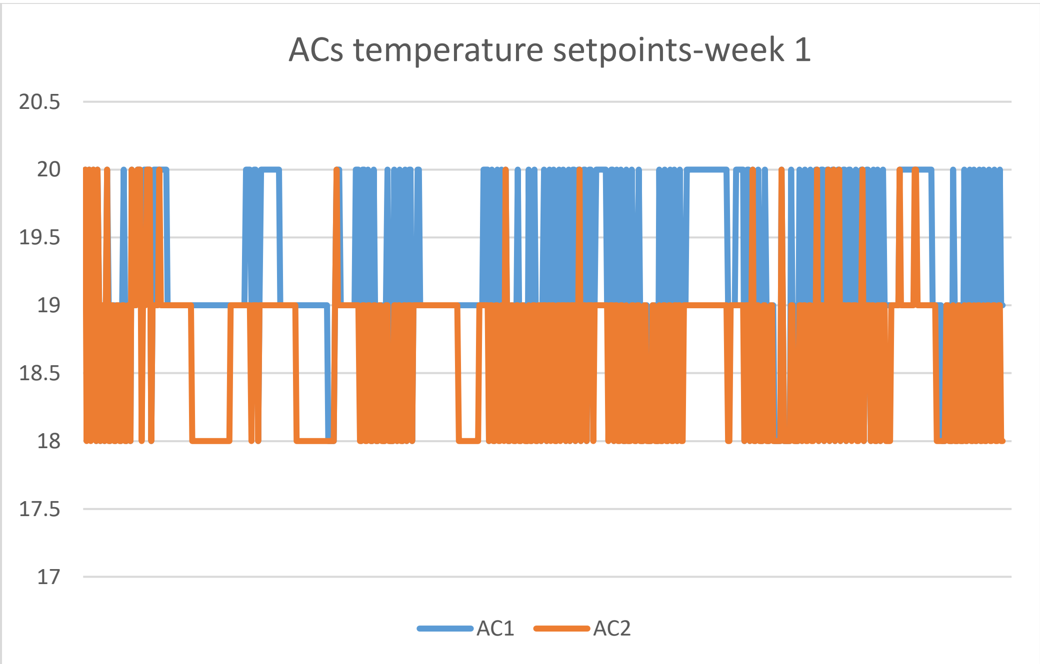


Figure 5-5 The historical data on the ACs' temperature setpoints in week 1

■ Other DC variables

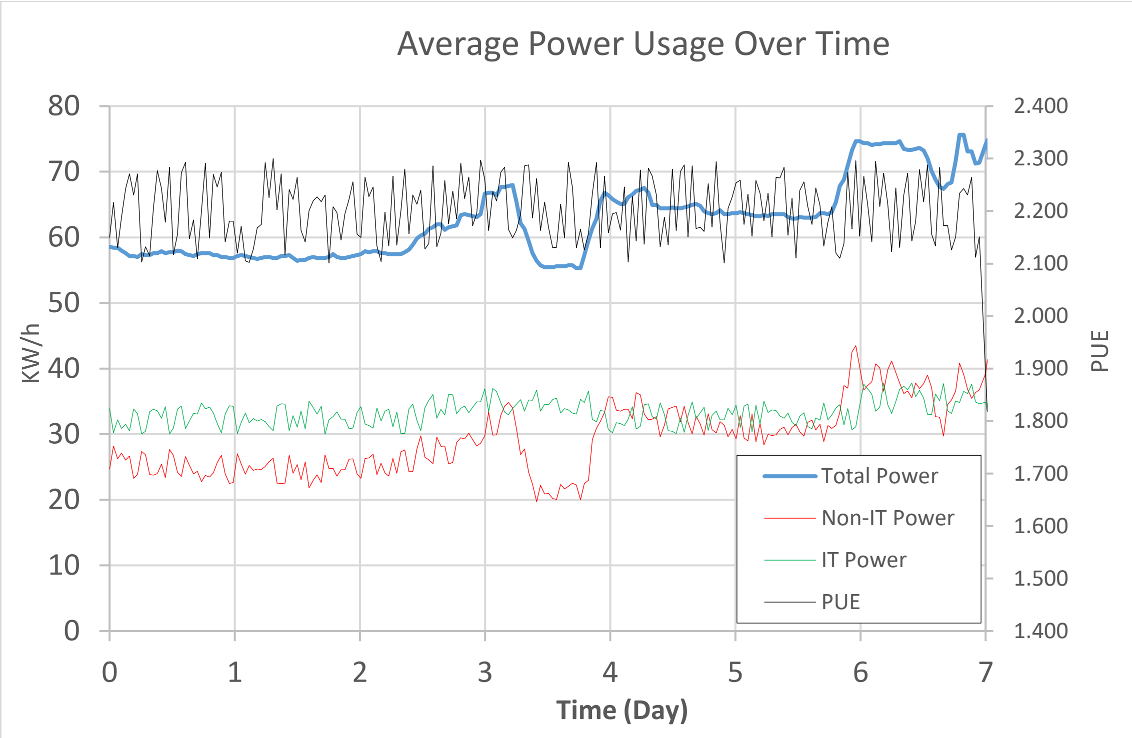


Figure 5-6 The historical data on power usage in Week 1

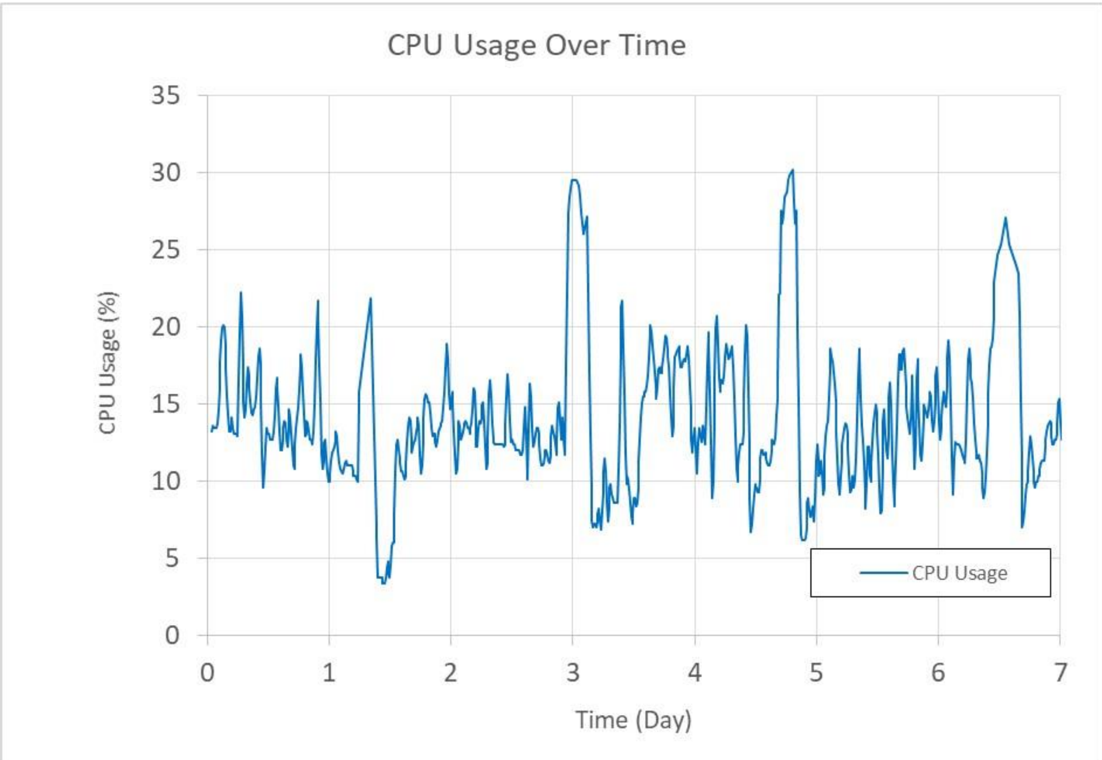


Figure 5-7 The historical CPU usage in Week 1

■ Descriptive statistics

Table 5-2 summarised the statistics of the historical data of experiment week I. The results show that before adopting the MPC control strategy, both ACs have been operated at a low temperature of around 18 to 19 degrees, and the low temperatures have been always kept for a long time. The ambient temperature of the server room was around 23.6 degrees. This is 4 degrees lower than the ASHRAE recommended maximum temperature boundary. PUE value was around 2.2, which indicated the DC has been operated relatively inefficiently.

During the experiment, an unexpected decline in non-IT power, coupled with a corresponding reduction in overall power consumption, was observed on the third and fourth days. This phenomenon was attributed to an unforeseen power outage. This anomaly has been acknowledged as an outlier, and as a result, the affected data has been deliberately omitted from the subsequent analysis.

Table 5-2 Descriptive statistics of the DC variables in experiment week 1

	T_OUTFLOW	T_OUTDOOR	T_AMBIENT	PUE	HG	CPU	AC2	AC1
Mean	19.16679	29.19552	23.65970	2.203755	9.052985	13.59510	18.66269	19.15224
Median	19.25000	29.00000	24.03270	2.207000	9.000000	13.07510	19.00000	19.00000
Maximum	20.25000	35.50000	25.00000	2.300000	11.00000	30.17150	20.00000	20.00000
Minimum	18.25000	24.00000	23.50000	2.100000	4.500000	3.404470	18.00000	18.00000
Std. Dev.	0.523618	2.732161	0.412640	0.058717	0.658813	4.373003	0.549177	0.670176
Skewness	-0.175490	0.319075	0.211739	-0.052855	-1.498150	1.019266	0.035139	-0.186834
Kurtosis	1.869689	2.467017	2.763941	1.726512	13.25361	5.582868	2.251979	2.202208
Jarque-Bera	39.10541	19.29898	6.562020	45.58641	3185.692	302.2486	15.75825	21.66611
Probability	0.000000	0.000064	0.037590	0.000000	0.000000	0.000000	0.000379	0.000020
Sum	12841.75	19561.00	16522.00	1476.516	6065.500	9108.715	12504.00	12832.00
Sum Sq. Dev.	183.4236	4993.887	113.9119	2.306504	290.3690	12793.39	201.7672	300.4716
Observations	670	670	670	670	670	670	670	670

5.3.2 Analytical results

Before building the VAR model, the analysis of whether the data required to meet the VAR conditions has to be conducted. Analysis including stability test, lag length selection for building VAR, the Granger causalities, impulse responses, and variance decompositions have been conducted after the VAR coefficients have been estimated. Additionally, the outdoor temperatures and CPU usage will be forecasted by univariate ARIMA models. These two variables, as well as the ACs setpoints, will be inserted into the VAR model as exogenous variables to forecast the PUE in future periods. To ensure the quantity of historical data, and present the data movement in more detail, the analytical part adopted 15 minutes of data resolution, as the raw data given in week 1. Raw data in experiment week 1 contains 672 sample points, after adjustment, the analysis contains 670 data points. As regards the tests that would be related to the model structure, hourly data with 167 data points for a week would be adopted, this is consistent with the following optimisation and MPC control data resolution.

a. Stability test

The VAR model requires all the endogenous variables to be stationary.

Table 5-3 shows the unit root test for the endogenous variables. The result shows that the unit root test rejected the endogenous variables that have a unit root. This means that the endogenous variables we have are stationary, and the system is stable. This condition is meet the requirements of building a VAR model.

Table 5-3 The unit root test results of endogenous variables

Group unit root test: Summary
Series: L_HG, L_PUE, L_T_AMBIENT, L_T_OUTFLOW

Sample: 1 670
Exogenous variables: Individual effects
Automatic selection of maximum lags
Automatic lag length selection based on SIC: 0 to 5
Newey-West automatic bandwidth selection and Bartlett kernel

Method	Statistic	Prob.**	Cross-sections	Obs
Null: Unit root (assumes common unit root process)				
Levin, Lin & Chu t*	-20.1367	0.0000	4	2664
Null: Unit root (assumes individual unit root process)				
Im, Pesaran and Shin W-stat	-21.4560	0.0000	4	2664
ADF - Fisher Chi-square	264.815	0.0000	4	2664
PP - Fisher Chi-square	600.381	0.0000	4	2676

** Probabilities for Fisher tests are computed using an asymptotic Chi -square distribution. All other tests assume asymptotic normality.

b. Lag length selection

It can be concluded from Table 5-4 that all of the criteria give the smallest number to the lag length of 1 and that the criterion tends to select 1 as the optimal lag length for the VAR model. This demonstrates how past information influences the dependent variable's behaviour after an hour.

Table 5-4 The lag length selection result and criteria

VAR Lag Order Selection Criteria
Endogenous variables: T_AMBIENT T_OUTFLOW PUE HG
Exogenous variables: C CPU AC1 AC2 T_OUTDOOR
Sample: 1 167
Included observations: 159

Lag	LogL	LR	FPE	AIC	SC	HQ
0	240.2719	NA	7.36e-07	-2.770716	-2.384690	-2.613955

1	290.3657	94.51670*	4.79e-07*	-3.199569*	-2.504723*	-2.917399*
2	302.1230	21.59199	5.06e-07	-3.146202	-2.142535	-2.738623
3	309.4342	13.05889	5.66e-07	-3.036908	-1.724420	-2.503921
4	316.5900	12.42148	6.34e-07	-2.925661	-1.304353	-2.267265
5	332.0295	26.02374	6.41e-07	-2.918610	-0.988481	-2.134805
6	344.8833	21.01876	6.71e-07	-2.879035	-0.640086	-1.969821
7	350.1073	8.279567	7.75e-07	-2.743488	-0.195718	-1.708866
8	357.3503	11.11503	8.74e-07	-2.633337	0.223253	-1.473306

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

c. Granger causality test

Table 5-5 demonstrated a significant result of the Granger causality test rejecting the null hypothesis that the variables do not have a Granger causality relationship. The conclusion is, that the past value of one variable can affect the future value of another, which means that the DC endogenous variables have interrelated causality.

Table 5-5 Granger causality test result

Pairwise Granger Causality Tests
Sample: 1 670
Lags: 2

Null Hypothesis:	Obs	F-Statistic	Prob.
PUE does not Granger Cause HG	668	1.03054	0.0357
HG does not Granger Cause PUE		1.19824	0.0302
T_AMBIENT does not Granger Cause HG	668	92.0593	5.E-36
HG does not Granger Cause T_AMBIENT		73.5063	1.E-29
T_OUTFLOW does not Granger Cause HG	668	154.422	9.E-56
HG does not Granger Cause T_OUTFLOW		23.7933	1.E-10
T_AMBIENT does not Granger Cause PUE	668	2.62599	0.0731
PUE does not Granger Cause T_AMBIENT		2.56248	0.0779
T_OUTFLOW does not Granger Cause PUE	668	1.38762	0.0254
PUE does not Granger Cause T_OUTFLOW		1.22143	0.0295
T_OUTFLOW does not Granger Cause T_AMBIENT	668	97.6517	7.E-38
T_AMBIENT does not Granger Cause T_OUTFLOW		42.0526	6.E-18

d. Impulse response functions

Figure 5-8 shows the impulse responses of each variable. If the variable is sensitive to another variable's shock, it will show an obvious trend in the graph.

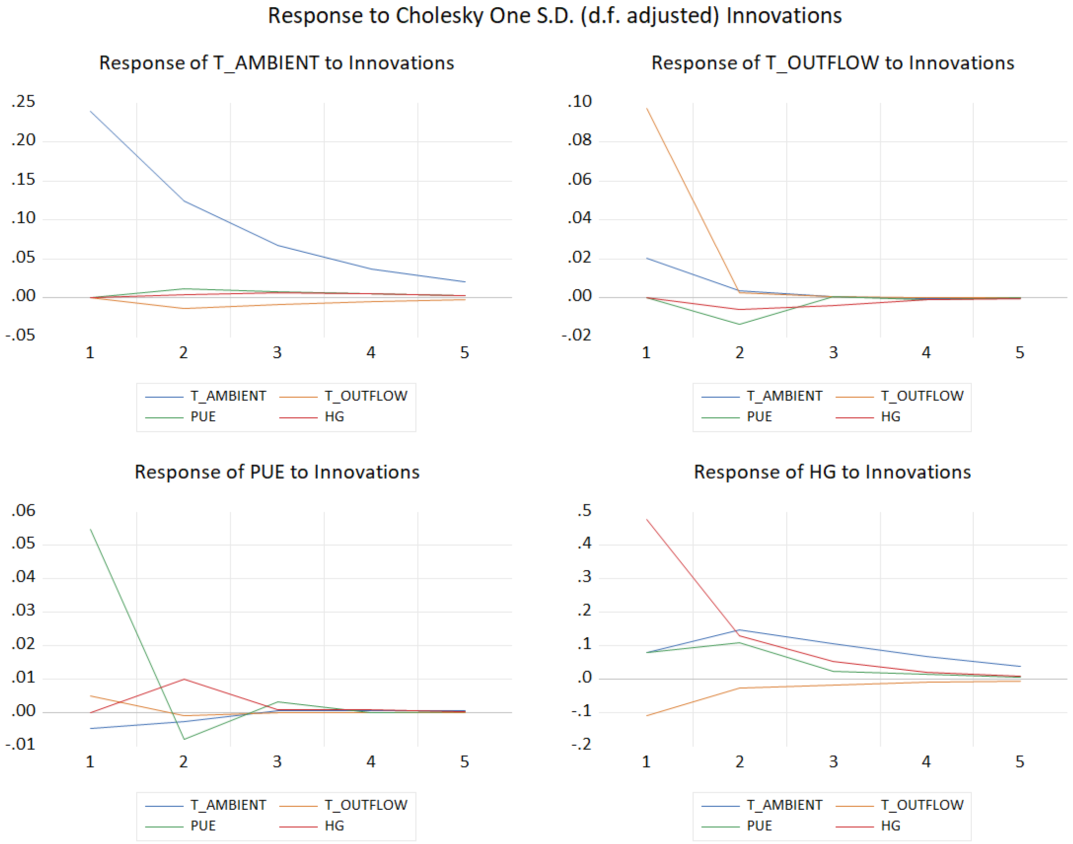


Figure 5-8 The impulse response of the endogenous variables

If we give each variable one standard deviation of shock from another variable, the trend that this variable will respond to is shown in Figure 5-8. These give us the information that the variables mutually affect each other, and the responses follow certain trends.

e. Variance decomposition

Figure 5-9 shows the variance decomposition of each endogenous variable. Variance decomposition indicates how the percentage changes of one variable to the other variables' shock. From the result, it can be seen that the DC variables mutually affect each other and have responded to the given shock from each other. After the above historical data analysis, we can build a VAR model with 4 endogenous variables, 4 exogenous variables, and a lag length of 1 at level (0 differential).

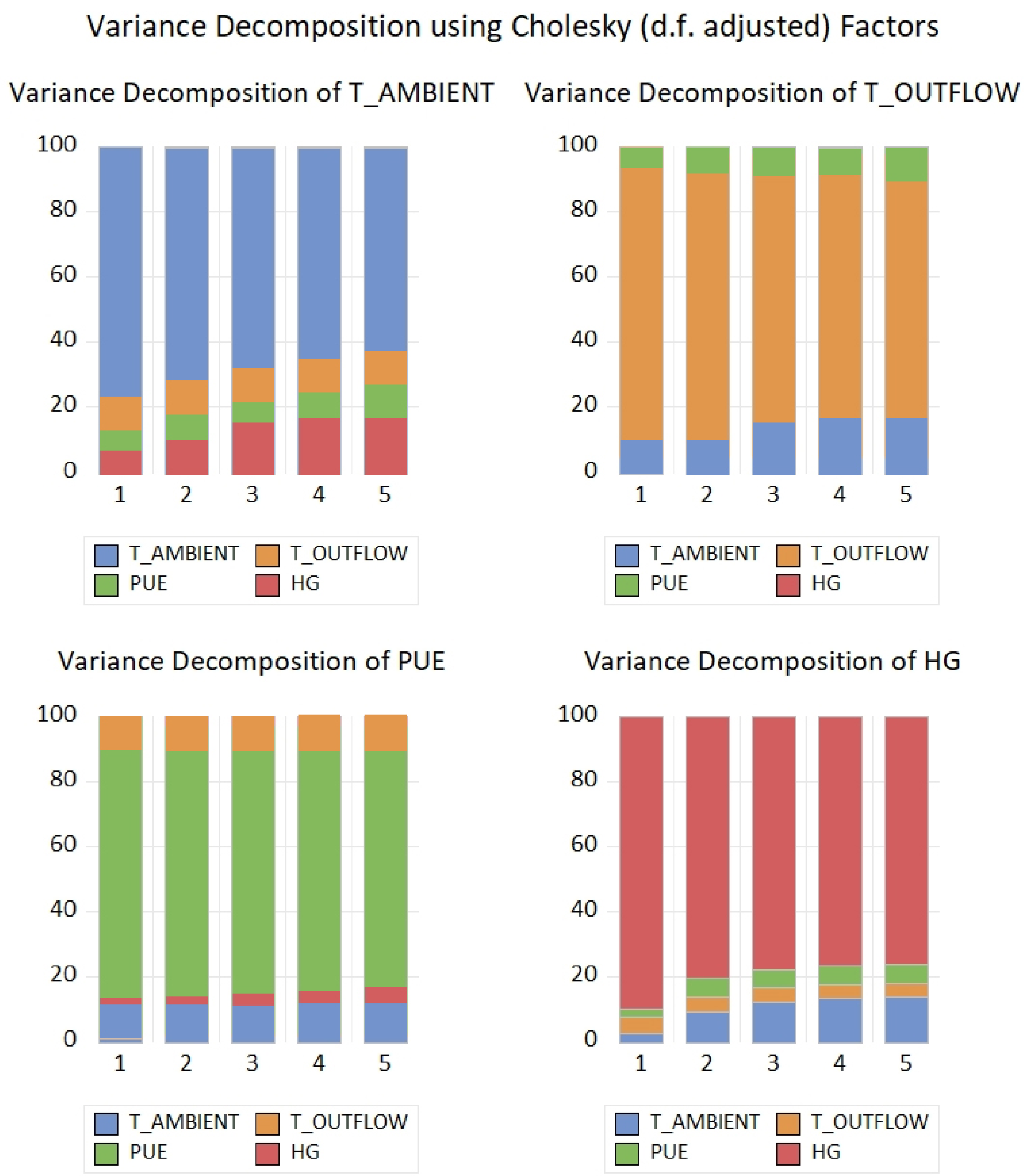


Figure 5-9 Variance decomposition of the endogenous variables

5.4 The prediction of exogenous variables

Exogenous variables in the VAR model have been considered as given, therefore separate forecasting models would be applied to forecast the exogenous variables, to be inserted into the VAR model to forecast PUE.

5.4.1 Prediction of the outdoor temperature $T^{outdoor}$

The exogenous variable $T^{outdoor}$ must be given when predicting the PUE value by the VAR model. ARIMA model will be adopted to predict the value of the CPU.

Table 5-6 demonstrates the correlogram of the $T^{outdoor}$ series. The ACF and PACF graphs illustrate the CPU following an AR process. ACF graph gradually declined and PACF cut-off at two spikes indicates that AR(2) model can be used to forecast future $T^{outdoor}$s. The AR(2) model of the outdoor temperature is shown in Table 5-7 The AR(2) model of outdoor .

Table 5-6 Correlogram of OT

Sample: 1 167
Included observations: 167

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.955	0.955	154.93	0.000
		2	0.873	-0.432	285.28	0.000
		3	0.773	-0.107	388.17	0.000
		4	0.663	-0.095	464.26	0.000
		5	0.548	-0.065	516.53	0.000
		6	0.440	0.053	550.40	0.000
		7	0.342	-0.006	570.99	0.000
		8	0.254	-0.020	582.46	0.000
		9	0.182	0.044	588.37	0.000
		10	0.131	0.100	591.47	0.000
		11	0.100	0.024	593.27	0.000
		12	0.084	0.024	594.55	0.000
		13	0.078	-0.037	595.66	0.000
		14	0.084	0.074	596.96	0.000
		15	0.102	0.083	598.91	0.000

Table 5-7 The AR(2) model of outdoor temperature

Dependent Variable: T_OUTDOOR
Method: ARMA Maximum Likelihood (OPG - BHHH)
Sample: 1 167
Included observations: 167
Convergence achieved after 12 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	29.07331	0.842731	34.49893	0.0000
AR(1)	1.391506	0.062856	22.13796	0.0000
AR(2)	-0.451324	0.060404	-7.471757	0.0000
SIGMASQ	0.463956	0.043718	10.61235	0.0000

R-squared	0.937497	Mean dependent var	29.20659
Adjusted R-squared	0.936346	S.D. dependent var	2.732692
S.E. of regression	0.689450	Akaike info criterion	2.135612
Sum squared resid	77.48060	Schwarz criterion	2.210295
Log likelihood	-174.3236	Hannan-Quinn criter.	2.165924
F-statistic	814.9531	Durbin-Watson stat	2.093974
Prob(F-statistic)	0.000000		

Inverted AR Roots	.88	.51

Figure 5-10 demonstrates $T^{outdoor}$ as forecasted by the AR(2) model. The forecasted value has been compared with the actual measured value in the graph. In the VAR prediction, the OT would be used as the exogenous variable to forecast the PUE value.

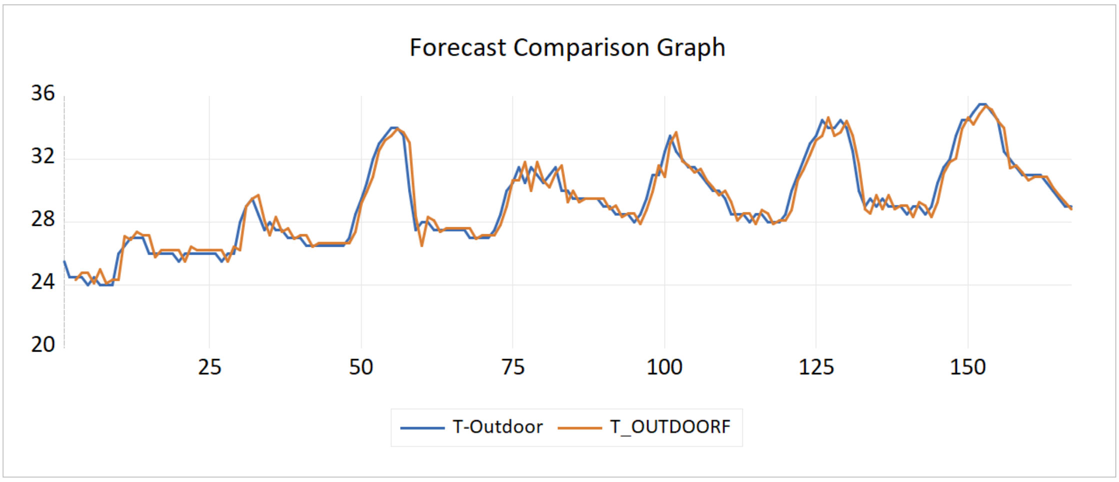


Figure 5-10 The prediction of outdoor temperature

5.4.2 Prediction of Workload

The exogenous variable workload (CPU usage) has to be given when predicting the PUE value by the VAR model. ARIMA model will be adopted to predict the value of the CPU usage.

Table 5-8 The correlogram of CPU usage. Table 5-8 demonstrates the correlogram of the CPU series. The gradually downward trend ACF graph and a PACF graph with two significant spikes illustrate the CPU following an AR process. An AR(2) model has been adopted to model and forecasts future

CPUs, the model is shown in Table 5-9.

Table 5-8 The correlogram of CPU usage

Sample: 1 167
Included observations: 167

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.631	0.631	67.793	0.000
		2	0.289	-0.182	82.120	0.000
		3	0.016	-0.146	82.165	0.000
		4	-0.034	0.120	82.359	0.000
		5	-0.089	-0.128	83.734	0.000
		6	-0.095	-0.015	85.302	0.000
		7	-0.102	-0.013	87.127	0.000
		8	-0.077	-0.022	88.181	0.000
		9	0.031	0.153	88.356	0.000
		10	0.120	0.029	90.924	0.000
		11	0.122	-0.047	93.638	0.000
		12	-0.048	-0.207	94.064	0.000
		13	-0.118	0.060	96.610	0.000
		14	-0.085	0.083	97.948	0.000
		15	0.001	-0.005	97.948	0.000
		16	0.050	0.053	98.422	0.000
		17	0.067	0.010	99.269	0.000
		18	0.003	-0.112	99.270	0.000
		19	-0.011	0.049	99.292	0.000
		20	0.005	-0.010	99.297	0.000

Table 5-9 The AR(2) model of CPU usage

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.54856	0.656119	20.64954	0.0000
AR(1)	1.409435	0.018057	78.05353	0.0000
AR(2)	-0.501225	0.017972	-27.88964	0.0000
SIGMASQ	1.690801	0.029732	56.86864	0.0000

R-squared	0.911451	Mean dependent var	13.59510
Adjusted R-squared	0.911053	S.D. dependent var	4.373003
S.E. of regression	1.304207	Akaike info criterion	3.379066
Sum squared resid	1132.836	Schwarz criterion	3.405975
Log likelihood	-1127.987	Hannan-Quinn criter.	3.389489
F-statistic	2285.099	Durbin-Watson stat	2.030795
Prob(F-statistic)	0.000000		
Inverted AR Roots	.70+.07i	.70-.07i	

Figure 5-11 is the CPU usage forecasted by AR(2) model. The forecasted value has been compared with the actual measured value in the graph. In the VAR prediction, the OT would be used as the exogenous variable to forecast the PUE value.

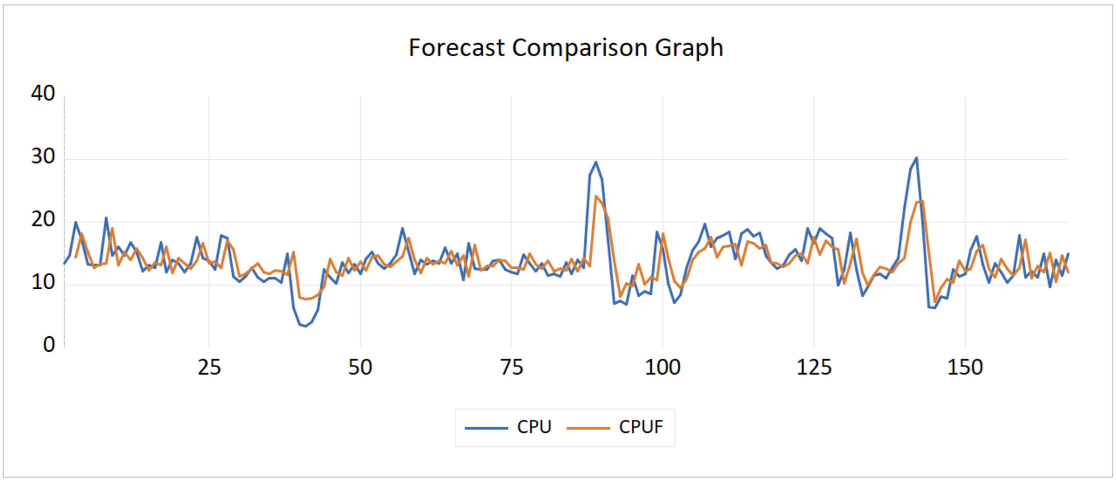


Figure 5-11 The prediction of CPU usage

5.5 The VAR model and an open-loop VAR-PSO optimisation

5.5.1 The VAR model structure

VAR modelled each variable as a dependent variable and depends on other variables' historical behaviours. Here we pick out the part where PUE act as the dependent variable and depends on other variables. The model adopted the time lag of one at the level. The model structure and parameters are shown in Table 5-10.

Table 5-10 The structure of the VAR model

Vector Autoregression Estimates
Sample (adjusted): 2 167
Included observations: 166 after adjustments
Standard errors in ( ) & t-statistics in [ ]

	T_AMBIENT	T_OUTFLOW	PUE	HG
T_AMBIENT(-1)	0.533443	0.011664	-0.023619	0.558555
	(0.07479)	(0.03099)	(0.01727)	(0.15725)
	[7.13242]	[0.37631]	[-1.36733]	[3.55212]
T_OUTFLOW(-1)	-0.145877	0.024865	0.021021	-0.053549
	(0.06290)	(0.02607)	(0.01453)	(0.13224)
	[-2.31921]	[0.95390]	[1.44700]	[-0.40492]
PUE(-1)	0.202065	-0.233020	-0.175958	1.569453
	(0.34243)	(0.14191)	(0.07909)	(0.71995)
	[0.59009]	[-1.64203]	[-2.22483]	[2.17995]
HG(-1)	0.008628	-0.012280	0.020554	0.268192
	(0.03677)	(0.01524)	(0.00849)	(0.07732)
	[0.23463]	[-0.80574]	[2.41998]	[3.46871]
C	6.500060	2.022553	2.381807	1.467307
	(1.41445)	(0.58617)	(0.32668)	(2.97383)
	[4.59546]	[3.45043]	[7.29085]	[0.49341]
CPU	0.006009	0.000224	-0.000848	0.017864
	(0.00453)	(0.00188)	(0.00105)	(0.00952)
	[1.32651]	[0.11936]	[-0.81050]	[1.87572]
AC1	0.246315	0.471300	0.005062	-0.255471
	(0.04007)	(0.01661)	(0.00926)	(0.08425)

	[6.14654]	[28.3791]	[0.54692]	[-3.03216]
AC2	0.098895	0.418820	0.007940	-0.367298
	(0.05019)	(0.02080)	(0.01159)	(0.10552)
	[1.97047]	[20.1366]	[0.68496]	[-3.48087]
T_OUTDOOR	0.022007	0.005404	-0.001100	0.015617
	(0.00750)	(0.00311)	(0.00173)	(0.01577)
	[2.93475]	[1.73900]	[-0.63513]	[0.99057]
R-squared	0.691963	0.970349	0.668377	0.457281
Adj. R-squared	0.676266	0.968838	0.020906	0.429626
Sum sq. resids	9.018443	1.548843	0.481074	39.86453
S.E. equation	0.239671	0.099324	0.055355	0.503899
F-statistic	44.08480	642.2307	1.440392	16.53549
Log likelihood	6.211634	152.4380	249.4852	-117.1442
Akaike AIC	0.033595	-1.728169	-2.897412	1.519810
Schwarz SC	0.202317	-1.559446	-2.728689	1.688532
Mean dependent	24.66867	19.15060	2.201946	9.069277
S.D. dependent	0.421233	0.562650	0.055943	0.667212

Determinant resid covariance (dof adj.)	3.76E-07
Determinant resid covariance	3.01E-07
Log likelihood	304.2388
Akaike information criterion	-3.231792
Schwarz criterion	-2.556904
Number of coefficients	36

The above VAR model will be used as the prediction model to predict the PUE value for the future N periods. This prediction model will be embedded into the fitness function in response to the changes in the ACs' temperature setpoints, to optimise the PUE by minimising the future gap between PUE forecasted and reference values.

5.5.2 An open-loop optimisation

An open-loop optimisation is a form of optimisation where the system being optimised is not continuously monitored during the optimisation process, and the optimisation is solely based on a predetermined model or set of inputs. To test the feasibility of the PSO solver, an open-loop optimisation based on VAR has been conducted without applying any feedback control strategy. Figure

5-12 illustrates the 24-hour PUE value forecasting before and after optimization for the initial loop. The pre-optimization curve represents the forecasted PUE using the initial AC temperature settings. Simultaneously, PSO optimisation has been applied to determine the optimal combinations of AC temperatures at time 0, aiming to achieve optimal PUE values for the subsequent 24 hours (depicted as the post-optimization PUE curve).The graph indicates that prior to optimisation, the forecasted PUE values for the next 24 hours ranged from 2.2 to 2.4. After optimisation, this range notably decreased to 1.8 to 2.0. Additionally, there's an observable upward trend in the optimised PUE, suggesting that the performance under open-loop control diminishes over time. This trend could be attributed to the decreased predictive accuracy of the VAR model over longer time horizons, highlighting the necessity for a closed-loop control with a shorter control horizon to ensure the degree of optimal.

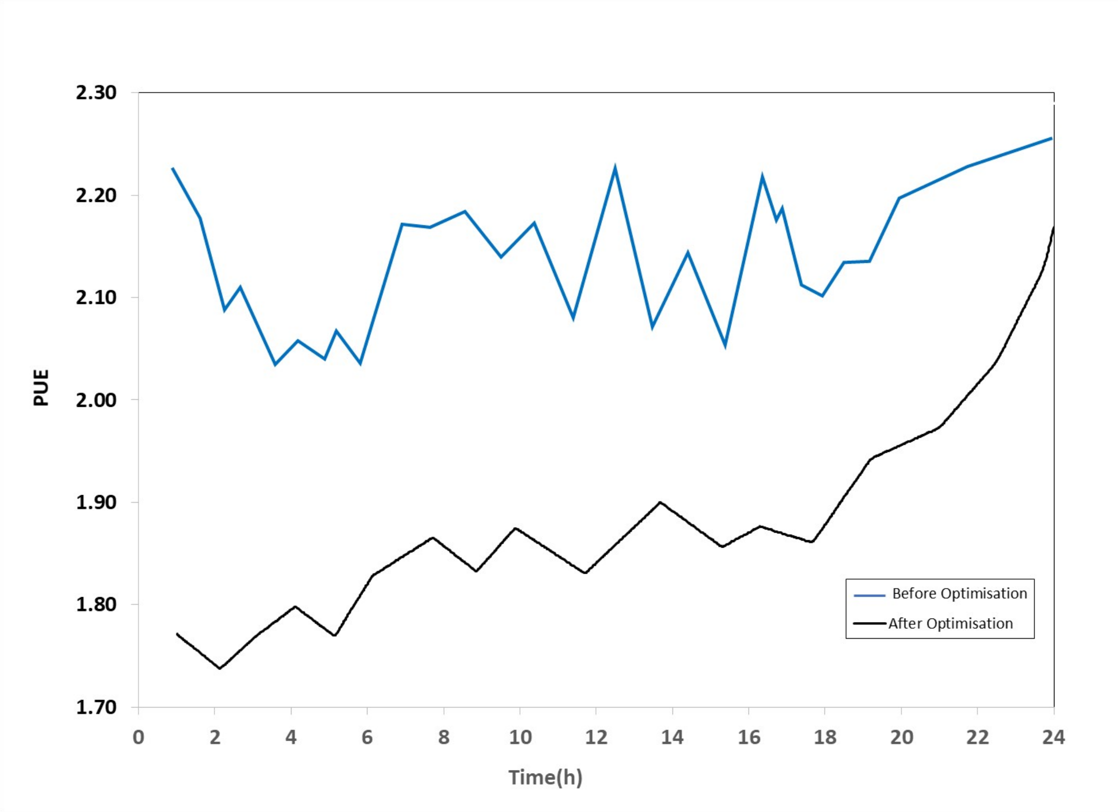


Figure 5-12 An open-loop control forecast for 24 hours

Table 5-11 summarised the statistic of the open-loop control, the result indicated a significant reduction in the value of the objective function and average PUE.

Table 5-11Summary statistics of one open-loop control

	Value of objective function	**Average PUE**
Before optimisation	11.7701	2.20
After optimisation	4.4308	1.86

5.6 The closed-loop MPC controlling

A closed-loop optimisation is a method of optimisation where the system under consideration is continuously monitored during the optimisation process, and the optimisation is modified based on real-time feedback from the system. In the previous section, the optimal ACs temperature setpoints have been implemented and kept constant for the whole period as an open-loop. However, it appears to be an upward trend. It can be inferred that keeping the ACs setpoints as the constant may be inefficient at the end of the control horizon. This is due to the decreased prediction performance of the prediction model at the end of the control horizon may cause a higher error for the control. Additionally, the air conditioner may work inefficiently at the end of the horizon which will cause PUE to increase. Therefore, in this section, MPC control will be introduced. The MPC keep the constant temperature setpoints on a shorter horizon, to ensure the optimal PUE in the long term.

Figure 5-13 demonstrates the result of the optimisation following the dynamic control strategy MPC. The upper part of the graph shows the predicted results of PUE values based on the optimal control strategies for four control loops. The grey curve on the upper graph is the actual PUE value under MPC control. The lower part of the graph shows the actual control actions on the cooling devices (the optimal ACs' temperature setpoints). It has been achieved by operating the suggested ACs' optimal setpoints at each control loop. At the start point of each loop, the optimisation model optimised the PUE value based on the prediction of future 24-hour performance and implement the suggested AC temperature setpoints recursively on an hourly basis.

At the period $k=0$ in the prediction horizon 1, the PSO optimiser optimised the ACs temperature setpoints by minimising $\sum_{i=1}^{24}[PUE_{0+i}-PUE_{ref}]^2$. The blue curve on the upper part of the graph shows the forecasting result based on optimisation. It can be inferred from the graph that there is an upward trend in 24 hours. The range is from 1.84 to 2.07, the first 6 periods are relatively low. Therefore, instead of keeping the optimal ACs' temperature setting to be constant for 24 hours, the optimal AC temperature settings have been kept on the ACs for less than 6 hours (The response of PUE shows as the red dot curve). To reduce the control error, the actual implementation takes just one hour at each control horizon. Then the time window moves to $t=k+1$. At prediction horizon 2, the PSO optimiser optimised the ACs temperature setpoints by minimising $\sum_{i=1}^{24}[PUE_{1+i}-PUE_{ref}]^2$. The orange curve on the upper part of the graph shows the forecasting result based on optimisation. Similarly, an upward trend has been shown in the graph after a few periods of prediction, the PUE value has been reduced to 2.01 at the end of the prediction horizon. Although PUE has been reduced compared to the first prediction horizon, to ensure the actual PUE is optimal and avoid the error due to the long-term prediction, the control actions have been implemented on the AC for just an hour. So far the control actions have been implemented for two hours, from period 1 to 2, and the corresponding PUE value has been shown as the red dot curve.

The same is for the prediction horizons 3 and 4, shown as the red dot in the graph. By applying the MPC control strategies recursively, the actual PUE value can be maintained at a relatively low range. As the grey curve that is shown in the figure.

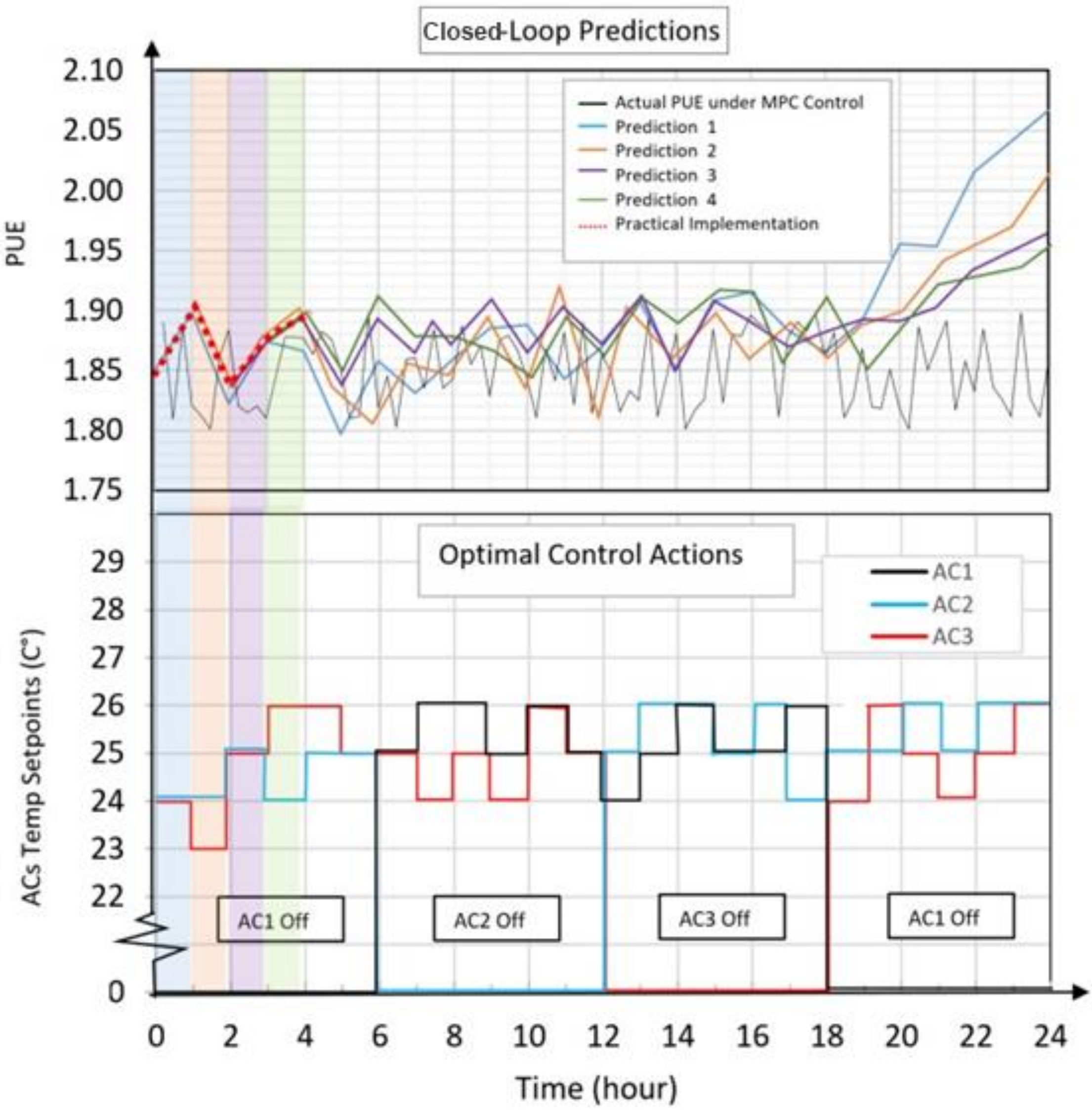


Figure 5-13 One-day closed-loop MPC control

5.6.1 The Performance Comparison experiment

(1) Comparison of different control intervals

Figure 5-14 compares the implementation of the MPC strategy on three different frequencies (hourly, 3-hourly, and 6-hourly). When comparing the mean of optimised PUEs of different implementation strategies, the PUE from implementing the strategy on an hourly frequency is more optimal than the PUE from implementing the strategy on a 3 and 6-hour frequency. It is possible to conclude that frequent and close monitoring and control can improve energy efficiency.

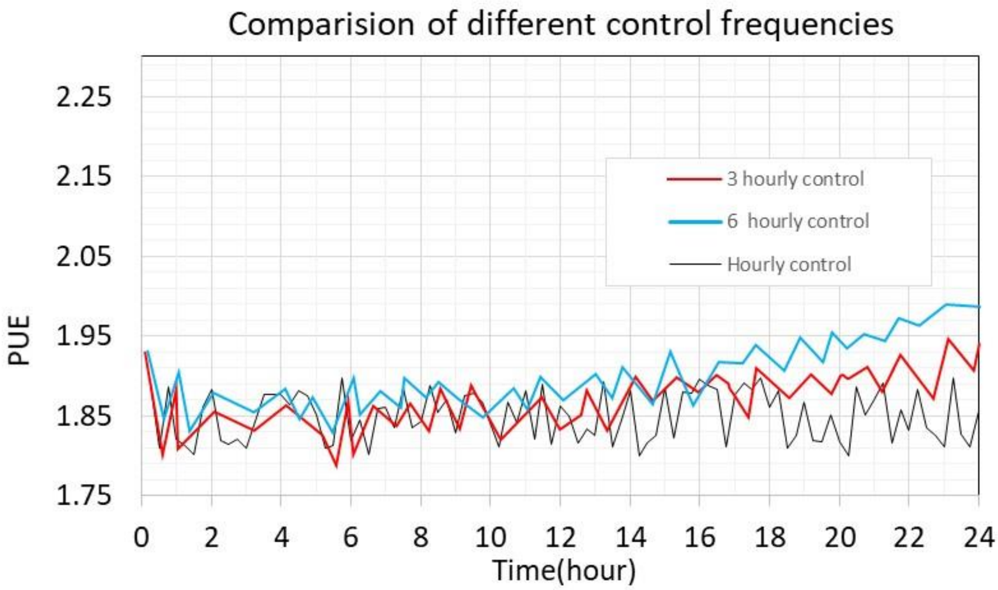


Figure 5-14 Optimised PUE with different implementation frequency

(2) The performance comparison of PSO and GA

As the characteristics of PSO and GA that have been summarised in the literature review section, PSO and GA are both heuristic optimisation algorithms that imitate an individual's adaptability to populations based on their natural characteristics, and because they both solve problems through the search space, their performances have been compared on various problems.

In this section, the same PUE optimisation problems have been solved with GA. To achieve this, the same algorithm has been applied with the same number of maximum iterations and population size. The performances of PSO and GA by comparing :(1) The mean of the fitness values that did not violate the constraint. Since it is a minimisation problem, so the smaller fitness value means a better performance of the minimisation. (2) The standard deviation of the fitness values to the mean. The smaller standard deviation means the optimisation system is more stable. (3) The running time of both algorithms. The shorter the better.

In order to identify the solutions that do not adhere to the constraints, a penalty of 1,000 has been applied to the fitness values that violated the constraints. Both PSO and GA have been run 50 times based on the initial settings. Table 5-12 shows the values of objective functions (fitness values), mean and standard deviations of the values that did not violate the constraints. Figure 5-15 plots the points of both fitness values for comparison. It can be concluded that with the same number of maximum iterations and population size, PSO has a better fitness value compared to GA. The standard deviation shows that among the fitness values, the PSO result is more stable than the GA result.

Table 5-12 The fitness values of PSO and GA

	Values of the objective function							
	[1]	[8]	[15]	[22]	[29]	[36]	Mean	Std.dv
PSO	4.717	5.125	5.157	4.234	4.019	3.384	4.364	0.598
	4.683	4.655	3.741	4.243	4.58	4.389		
	4.388	3.765	5.141	4.322	4.579	4.018		
	5.897	4.43	4.341	5.716	3.943	3.738		
	4.022	4.789	3.689	4.377	3.637	4.75		
	4.121	5.031	3.925	4.058	4.38	3.561		

	3.822	3.693	5.492	4.971	3.824	3.956		
	6.769	5.923	7.045	5.671	5.395	5.077		
	5.82	5.238	6.585	5.862	5.917	5.146		
	5.062	5.347	6.154	5.357	6.551	6.029		
GA	6.944	5.644	5.399	6.888	5.576	5.171	5.754	0.647
	4.894	5.976	5.898	5.883	4.918	5.046		
	4.638	5.276	4.965	5.538	6.731	5.737		
	6.713	5.851						

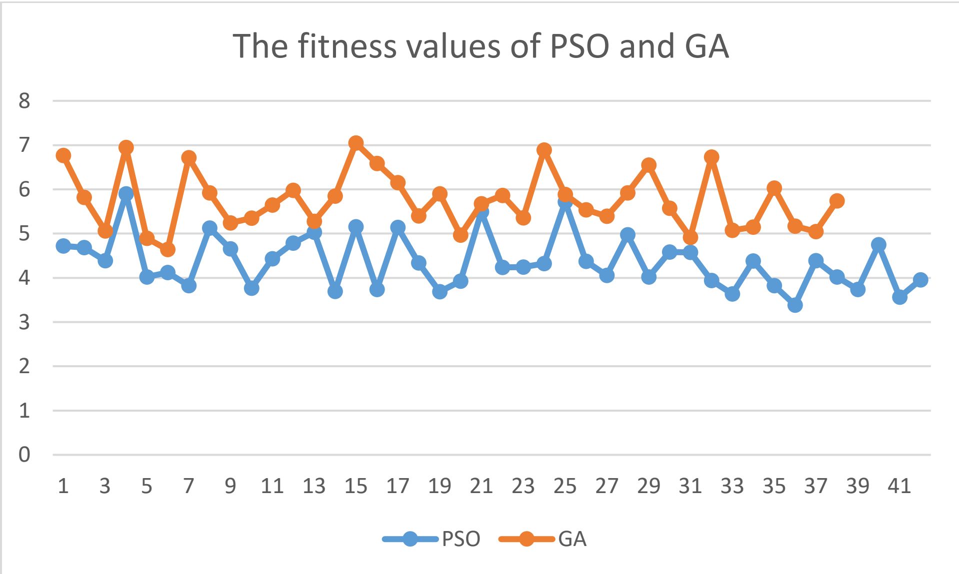


Figure 5-15 The fitness values of PSO and GA

The comparison of fitness values

Table 5-13 compared the running time of PSO and GA for one round. It can be concluded that with the same number of maximum iterations and population size, PSO has taken a shorter time to reach the optimal solution compared to GA. This can also be illustrated from the fitness values, PSO is more sensitive to identifying those solutions that violated the constraints (we extracted by adding the number 1,000), therefore narrowing the searching space to reduce the total time cost.

Overall, PSO optimisation outperforms the GA algorithm according to its fitness values, stabilities, and execution times.

Table 5-13 The running time of PSO and GA

	Running time
PSO	91.21 sec
GA	176.29 sec

5.7 Real-world Validation

The first part of the section introduces brief information on the validation in terms of the method and the location. Afterwards, three main components of validation have been presented in this section. As the performances of the system model VAR plays an important role in the prediction, which lays the foundation of the optimisation and system control, a validation of the prediction accuracy, has been provided.

The second part of the section explains the validation of the energy-saving potentials of the proposed approach, the experiment results from the previous chapter have been compared with the general energy consumption in the DC according to the information provided by the DC managers.

Additionally, the third part of the section verified the violence of environmental constraints. A detailed summary is presented to highlight the main findings.

5.7.1 Method of validation

MPC control strategy consists of optimisation and prediction, where the prediction function is embedded into the optimisation function. Therefore, validation should involve three aspects: The accuracy of the prediction model, the goodness of the optimisation, and the performance of the whole MPC control framework.

The validation has been conducted in the actual DC located in Thailand. After historical data collection and model adjustment, the validation has been conducted in the third week and the fourth week.

The validation indices include the accuracy of the prediction model, weekly percentage of energy saving, and percentage of PUE weekly reduction. By comparing the real data, the performances of the proposed approach can be validated in practice.

5.7.2 Validation of the prediction accuracy

The evaluation of VAR prediction is shown in Table 5-14. The Root Mean Squared Errors (RMSE) of four VAR endogenous variables range from 0.09 to 0.49, indicating that the prediction model performed well. Figure 5-16 to Figure 5-19 shows the comparison of forecasted and measured endogenous variables. The VAR model shows a good ability to handle multi-variable forecasting problems.

Table 5-14. The VAR forecast evaluation

Forecast Evaluation
Sample: 1 167
Included observations: 167

Variable	Inc. obs.	RMSE	MAE	MAPE	Theil
HG	167	0.490049	0.301708	3.320369	0.026964
PUE	167	0.053833	0.046380	2.106583	0.012222
T_AMBIENT	167	0.233084	0.190490	0.772288	0.004724
T_OUTFLOW	167	0.096594	0.068915	0.362521	0.002521

RMSE: Root Mean Square Error
MAE: Mean Absolute Error
MAPE: Mean Absolute Percentage Error
Theil: Theil inequality coefficient

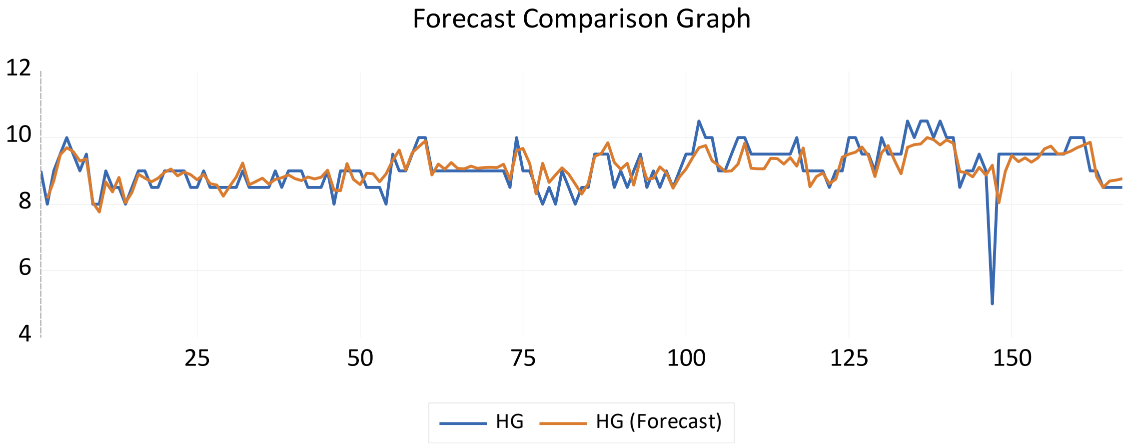


Figure 5-16 The forecast comparison of the heat generation by the servers

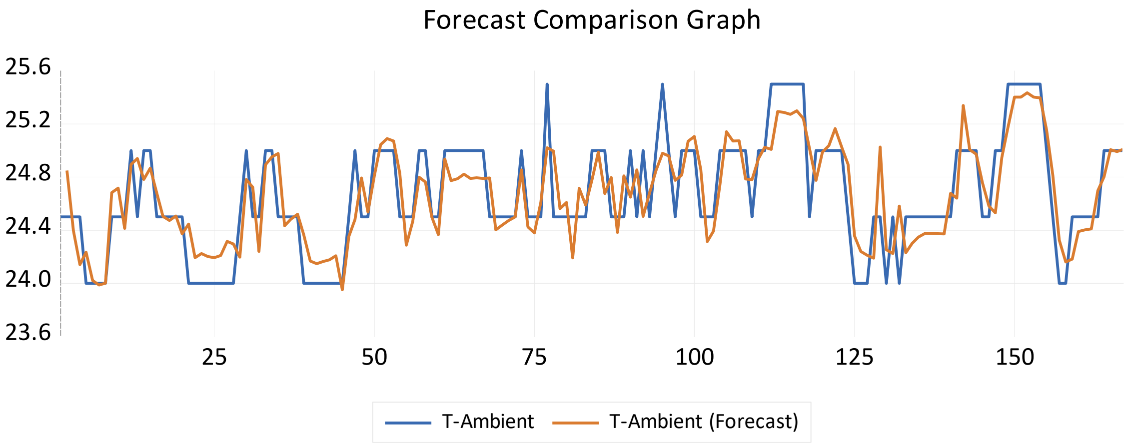


Figure 5-17 The forecast comparison of the ambient temperature in the server room

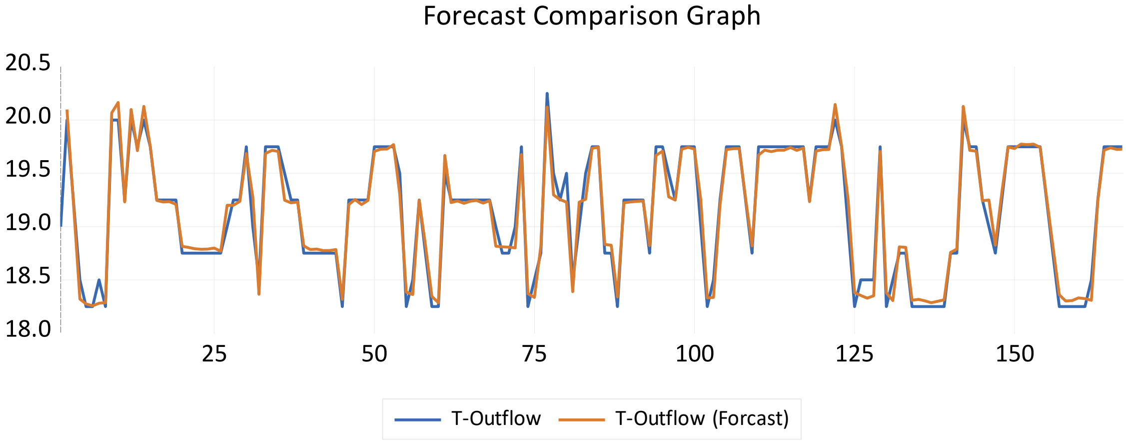


Figure 5-18 The forecast comparison of the outflow temperature of the AC

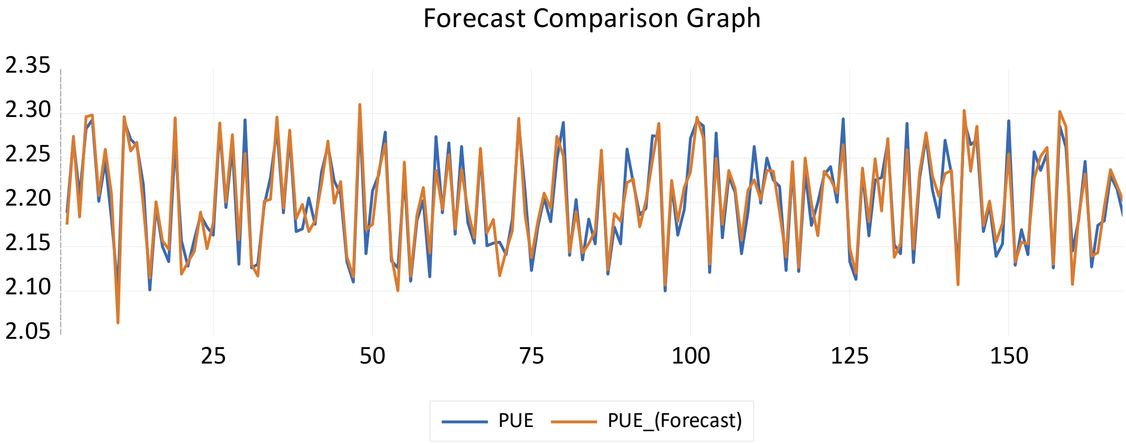


Figure 5-19 The forecast comparison of the PUE

5.7.3 Validation of the energy-saving performance

Figure 5-20 depicts the power consumption variables throughout the experiment. The graph shows that the total power is high in the first week without intervention about the server room nature. However, it began to fall dramatically on day 8, when the experiment started. It is primarily due to a significant reduction in Non-IT power consumption. Even after the second week, when the DC was asked to increase the IT load for the experiment, there was no increase in total power consumption, demonstrating that optimal AC temperature setpoints control is efficient.

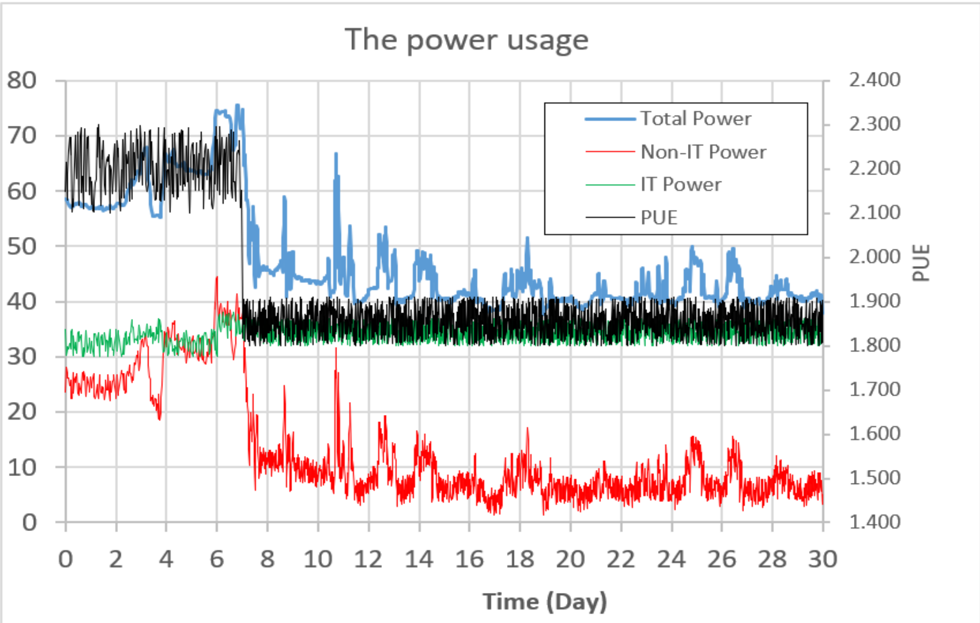


Figure 5-20 The changes in power usage variables

Figure 5-21 and Table 5-15 demonstrated the energy-saving potential in the target data centre by adopting the proposed MPC control strategy respectively. It can be concluded that controlling the AC setpoints using the MPC strategy can significantly reduce energy consumption in the DC and has a promising energy-saving potential.

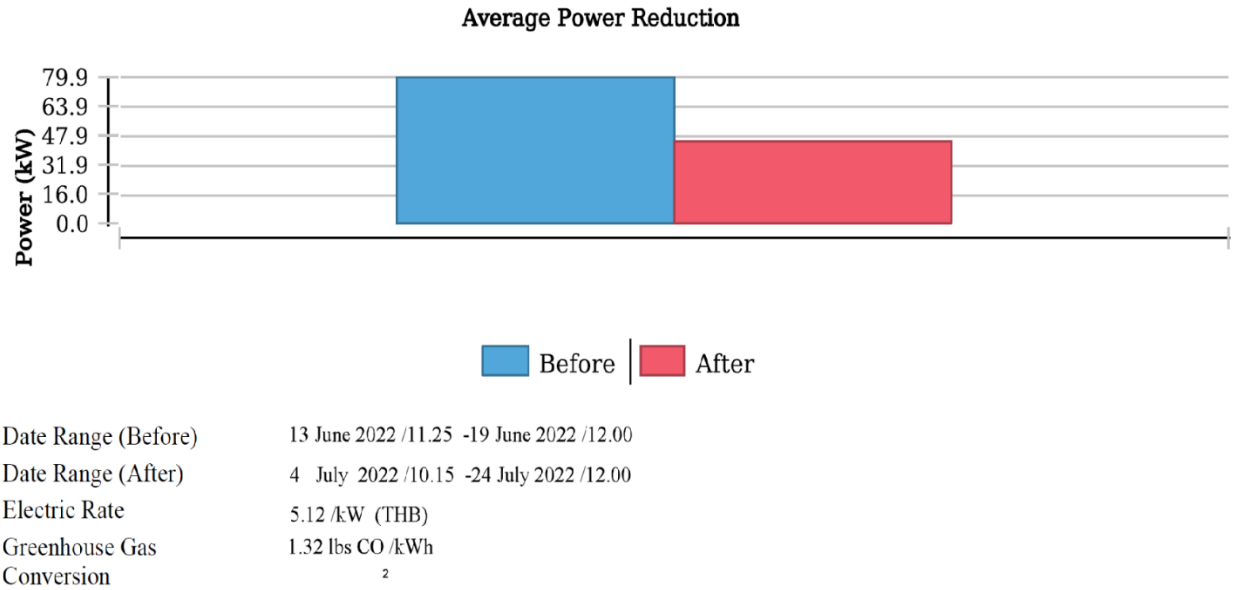


Figure 5-21 Average power before and after the experiment

Table 5-15 The energy-saving potential

Power				Annual Savings		
Before (kWh)	After (kWh)	Savings (kWh)	Reduction %	Energy savings kW	Cost savings (THB)	Greenhouse Gas Reduction lbs
63.51	42.86	20.65	32.5%	307,505	1,537,524	405,906

Table 5-16 summaries the descriptive statistics after implementing MPC strategies in the experiment from week 2 to week 4. The PUE mean value has decreased from 2.20 to 1.86. The server room ambient temperature has been increased by 1 to 2 degrees due to the increased ACs' temperature setpoints.

Table 5-16 Descriptive statistics of experiment Week 2 -Week 4

	T_OUTFLOW	T_OUTDOOR	T_AMBIENT	PUE	HG	CPU	AC2	AC1
Mean	22.28012	29.77654	26.07670	1.859797	8.945329	17.57359	21.77866	22.59713
Median	23.50000	29.50000	25.50000	1.862000	9.000000	16.58105	23.00000	24.00000
Maximum	24.50000	36.00000	28.00000	2.300000	13.00000	30.17150	24.00000	25.00000
Minimum	18.25000	24.00000	23.00000	1.800000	4.500000	3.404470	18.00000	18.00000
Std. Dev.	2.368723	2.478738	1.365139	0.116748	1.209664	6.000674	2.372696	2.621779
Skewness	-0.568551	0.240930	-0.003063	2.213539	0.414960	0.344680	-0.573126	-0.603050
Kurtosis	1.518326	2.464483	1.375815	6.742580	3.579858	2.205873	1.527139	1.564252
Jarque-Bera	273.8362	40.73897	207.0842	2638.064	80.46266	86.80967	273.4321	276.0101
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	41975.75	56099.00	49128.50	3571.681	16853.00	33108.64	41031.00	42573.00
Sum Sq. Dev.	10565.23	11569.42	3509.167	25.66561	2755.369	67803.23	10600.70	12943.22
Observations	1884	1884	1884	1884	1884	1884	1884	1884

5.7.4 Validation of the room and server temperature constraints

According to the ASHRAE guidelines, the server room temperature should be between 18°C and 27°C. Furthermore, the working temperatures of the servers should not be higher than 32°C or lower than 15°C. As shown in Figure 5-22 and Figure 5-23, the temperature from the front and back sides of the server rack temperature sensors, as well as the ambient temperature sensors, did not violate the boundary value during the experiment. This indicates that the server room's IT devices are operating under safe conditions.

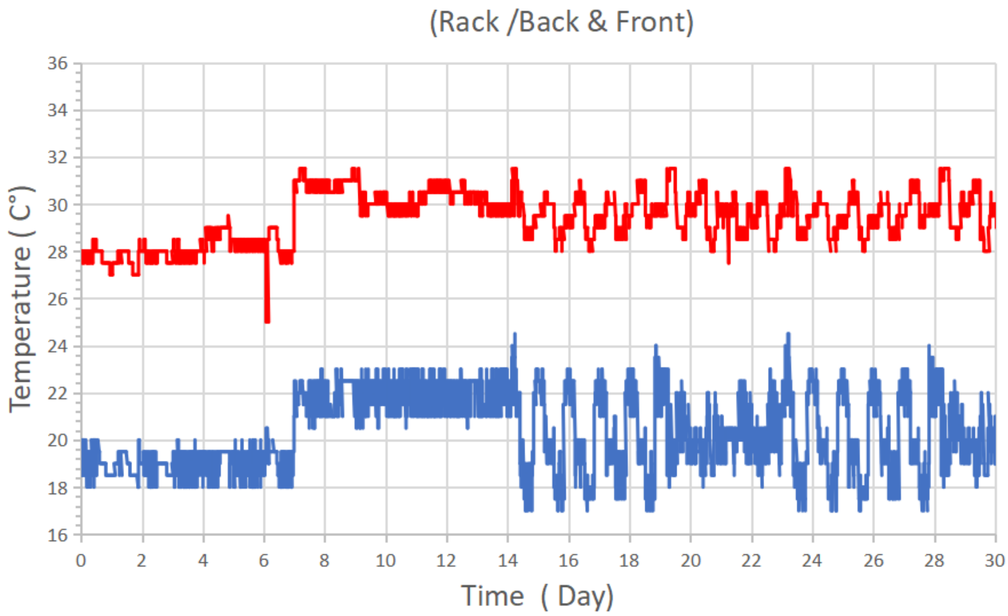


Figure 5-22 The rack front and back temperature during experiment

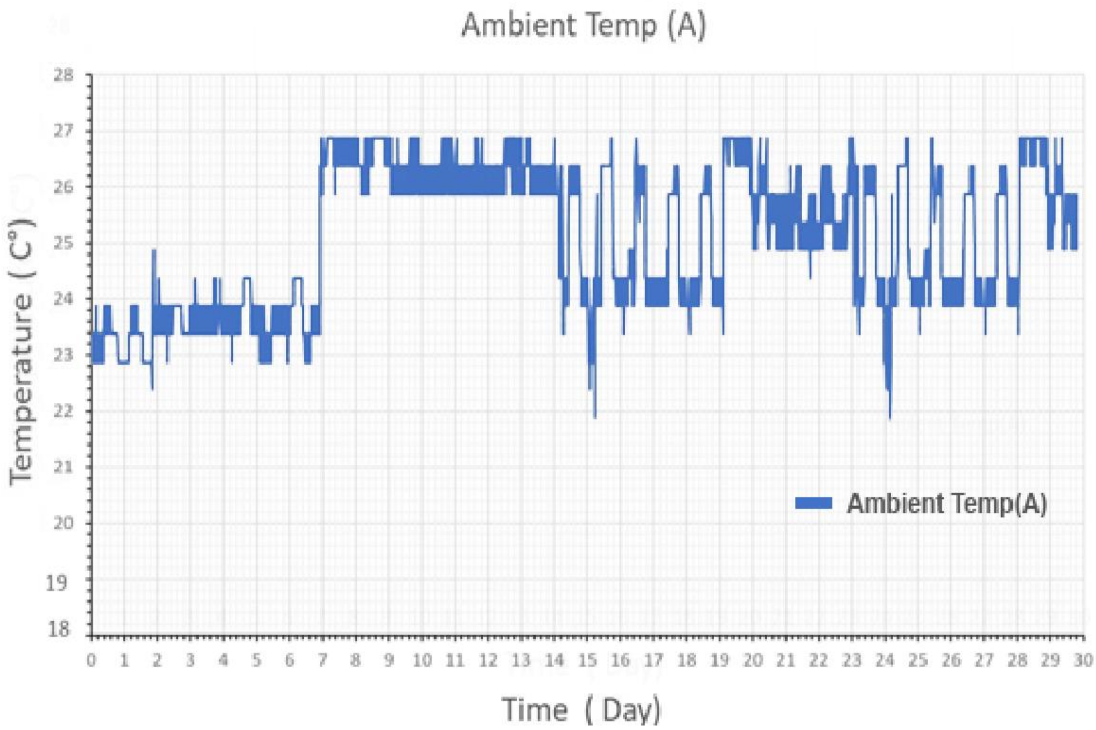


Figure 5-23 The ambient temperature during the experiment

CHAPTER 6

SIMULATION-BASED EVALUATION

This chapter presents the verification of the VAR-PSO-MPC approach in minimising the Power Usage Effectiveness (PUE) in data centres (DCs) using Computational Fluid Dynamics (CFD) simulation environment. The primary objective of this experiment is to test the performance of the VAR-PSO-MPC framework under different scenarios in the server room.

6.1Simulation purpose and overview

Testing a proposed approach in an actual DC in the real world can be a valuable way to validate its effectiveness and evaluate its practicality. However, there are several limitations to consider when conducting such tests:

1) Conducting tests in an actual DC can be expensive, particularly when it requires modifications to the current infrastructure. This may not be a viable option for smaller organizations or those with limited budgets due to the high costs involved.

2) Testing a new approach in a live DC can potentially lead to disruptions in ongoing operations, which could result in downtime or other issues. This can be particularly problematic in critical systems where any downtime can have severe consequences.

3) Limited scalability: Depending on the size and fix the configuration of the server room and devices, it may not be possible to test the approach at scale, which could limit the ability to draw meaningful conclusions.

4) Limited control: In a real-world DC, there may be factors beyond the experimenter's control that could impact the results of the test, such as changes in workload, network issues, or hardware failures.

5) Security and privacy: Testing a new approach in a live DC may raise security and privacy concerns, especially if it involves sensitive data or systems.

Given these limitations, it's often advisable to experiment in a simulated or controlled environment. This can help to identify potential issues and refine the approach while the actual data centre could not make it under some special or extreme conditions.

The purpose of the simulation is to address several key questions, including:

(1) How effective is the proposed VAR-PSO-MPC control strategy in reducing PUE?

(2) How quickly does the control strategy respond to changes in IT load and associated temperature increases?

(3) Does the control operation approach the upper and lower boundary values for room and server temperatures?

(4) How quickly does the proposed control strategy respond to extreme conditions?

(5) How well does the control framework adapt to changes in infrastructure, such as alterations to the server rack and air conditioner layouts?

To answer these questions, the simulation will be conducted using a baseline DC model followed by the application of the VAR-PSO-MPC framework under various operating conditions.

6.2 Simulation design

6.2.1 Simulation parameters

The effective design and management of DCs are critical to ensuring their smooth operation and maximum efficiency. A key aspect of this is the simulation of different parameters and conditions that may affect the performance of DCs. This section lists the simulation parameters that will be used to simulate the DC fluid dynamic and evaluate the performance of the control operations. The parameters include the size of the DC, the number and type of servers, the server layout, the cooling system, and the environmental conditions. Additionally, the locations and numbers of sensors installed will also be discussed.

(1) Size of the data centre

The size of the DC is an important factor that can impact its efficiency and effectiveness. In this experiment, the data centre size will be 54 square metres. This size is chosen based on the actual experiment DC size and the requirement to house 10 server racks while providing adequate space for movement and maintenance activities. A different size would have been ideal for future expansion of the experiment, however, due to the requirement of historical data, a similar DC size to the actual DC has been chosen for this experiment.

(2) Number and Type of Servers

In this experiment, 10 server racks with half of the servers on a high workload, and another half of the servers with a low workload have been adopted for the simulation. High-Performance Computing (HPC) servers are designed to handle complex and intensive workloads that require high-speed processing and storage capabilities. With the adoption of HPC servers, the workload can be simulated as that closely resembles real-world scenarios, making it easier to evaluate the performance of the DC. The use of HPC servers also allows for sufficient load balancing and redundancy, ensuring that the DC remains operational even in the event of a server failure.

(3) Server Layout

The layout of servers in a DC is another critical parameter that can affect its performance. In this experiment, the servers will be arranged in a hot aisle/cold aisle arrangement. This layout is a popular choice for DCs because it helps to optimise cooling and reduce energy consumption. In a hot aisle/cold aisle arrangement, servers are placed in alternating rows with cold air being supplied through the raised floor to the front of the servers (cold aisle) and hot air being exhausted from the back of the servers into the ceiling (hot aisle). This layout helps to prevent the mixing of hot and cold air, which can lead to inefficiencies and overheating.

(4) Cooling System

The cooling system used in a DC is crucial to maintaining optimal operating conditions. In this experiment, we will be using two air conditioners with underground airflow pipes. The air conditioners will be located in the area beside the server room and will supply cool air through the raised floor to the cold aisle. The hot air will be exhausted into the atmosphere through the ceiling. The use of underground airflow pipes helps to reduce energy consumption by minimising the distance that cool air needs to travel to reach the servers.

(5) Environmental Conditions

To ensure that the DC remains within the recommended operating conditions for its equipment, the environmental conditions in this experiment were referenced from the AHREA guideline (2018). As such, the ambient temperature boundary will be set at 17-28°C, which aligns with the actual experiment, and the relative humidity range will be 40-60%. These conditions have been determined to be optimal for the operation of high-performance computing servers and will help to maintain stable operating conditions within the DC. These environmental boundaries will serve as the basis for formulating the boundary conditions of the simulation in this experiment.

(6) Sensor Installation Points

The installation of sensors in a DC is essential to monitor and control its performance. In this experiment, we will be installing sensors in three locations, the locations and the number of the sensors will be displayed like location/number, as the following: ceiling/2, server racks/10, and AC outflow (air intake point)/1. The ceiling sensors will be used to monitor the overall temperature and humidity conditions within the DC. The server rack sensors will be used to monitor the temperature of the servers and detect any hotspots. These parameters are selected to reflect typical DC operating conditions and to allow for comparison with different operation scenarios.

6.2.2 Simulation steps

Computational Fluid Dynamics (CFD) is a popular technique used to simulate fluid flow in various applications, such as cooling systems in DCs. In this context, a CFD simulation can provide insights into the thermal performance of IT equipment and cooling systems, enabling engineers to optimise their design and operation. However, designing and running a CFD simulation can be a challenging task, involving several steps that require careful consideration. This section presents the crucial steps necessary for designing and executing a CFD simulation to model fluid flow in a DC cooling system. Moreover, it elaborates on how the simulation will validate the VAR-PSO-MPC control framework.

The following steps involve CFD simulation procedures:

1) Problem Definition: The first step is to define the problem, including the type of fluid flow, create the geometry of the system, and set the initial and boundary conditions.

2) Mesh Generation: The next step is to create a mesh that will discretise the geometry into small elements. The size and shape of the mesh will affect the accuracy of the results.

3) Selection of Fluid Model: The selection of a fluid model depends on the characteristics of the fluid being simulated. The most common models are laminar flow, turbulent flow, and multiphase flow.

4) Boundary Conditions: Boundary conditions are the constraints imposed on the fluid flow by the surrounding environment. These include inlet and outlet conditions, as well as any walls or solid objects in the system.

5) Solver Setup: Set up the CFD solver that is capable to govern the

equations of the fluid flow. The solver should be configured to accurately solve the problem defined in step 1.

6) Running the Simulation: Once the solver is set up, the simulation can be run. The output will be a set of data files containing temperature information about the fluid flow and also will give out the calculation result of the power usage based on the attributive and performance of the IT and cooling devices.

7) Post-Processing: The simulation must be post-processed to extract useful information. This includes creating visualisations of the fluid flow, as well as extracting numerical data for further analysis.

8) Validate the simulation: To ensure that the simulation results are accurate and reliable, Validation and verification tests will be performed. The simulation will be compared to the experimental data from the previous experiment, given that the simulation environment's size and layout are identical to that of the actual DC. Additionally, grid convergence studies and sensitivity analyses will be conducted to verify the accuracy of the simulation.

Once the CFD simulation environment is developed, it can be utilized to implement the MPC framework. The simulation takes the MPC-optimised setpoint temperature as input and produces an output in the form of a CSV formatted file. This output file contains the information required for the optimisation and control process to calculate the next optimised control input for the simulated environment. By utilizing this iterative process, the MPC algorithm adjusts the AC setpoint temperature to maintain the desired temperature conditions within the DC, at the same time, minimises PUE. This approach enables real-time control of the cooling system, ensuring that the

cooling capacity is always matched to the thermal load within the DC. Figure 6-1 demonstrates a processing graph of the proposed approach.

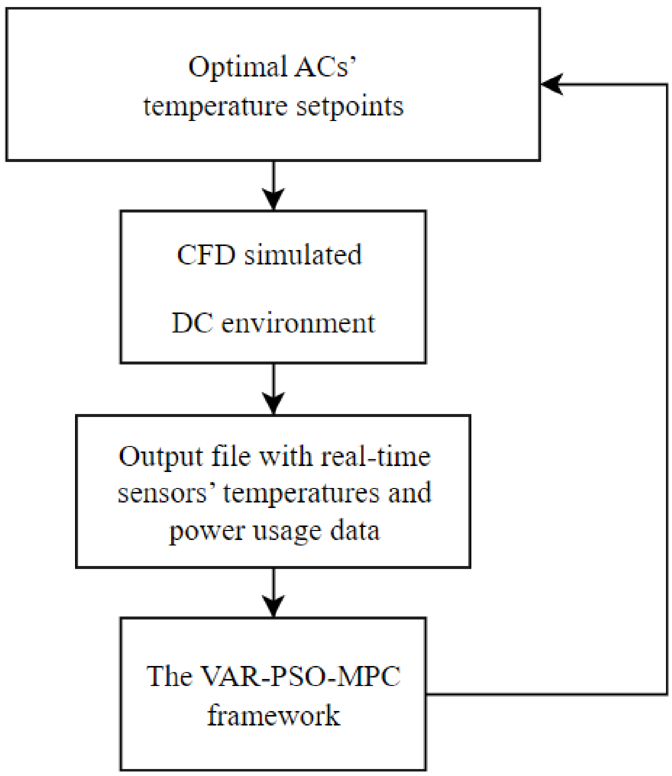


Figure 6-1 The process of validation under CFD environment

6.2.3 Performance Evaluating KPIs and Scenarios

6.2.3.1 KPIs

The following KPIs can be used to evaluate the performance:

(1) PUE: The PUE should be measured and compared under baseline

conditions and MPC control. The goal is to achieve a lower PUE with an MPC control framework.

(2) Temperature: The temperature of the server room should be monitored and compared under baseline conditions and MPC control. The goal is to achieve a more stable and optimal temperature with MPC control.

(3) Power Consumption: The power consumption of the IT equipment and air conditioning system should be monitored and compared under baseline conditions and MPC control. The goal is to achieve a lower power consumption with MPC control.

(4) System running time: Increasing the number of controlled variables: and. modifying the physical layout of the server room may have an impact on the system running time. A well-designed system should be able to adapt to these changes with minimal impact on its performance.

6.2.3.2 Scenarios

In addition to evaluating the performance of the MPC controller under different operating conditions, different scenarios can be simulated to test the robustness and flexibility of the controller. The scenarios that will be simulated in this experiment are as the following:

(1) Baseline: DC operating under typical conditions without the VAR-PSO-MPC framework applied.

(2) Server load variation: DC operating under variation of workload with the

MPC control applied to test the efficiency of the algorithm's response to changing workload conditions.

(3) Additional cooling units: The incorporation of extra cooling units aims to evaluate the adaptability of the system model and the performance of the MPC framework under conditions of increased cooling capacity, heightened cooling energy consumption, and variations in room layout.

(4) Additional server rack: Aims to assess the adaptability of the system model and evaluate its capacity to handle the significant increase in server workload and the subsequent rise in server and room temperatures, as well as variations in room layout.

(5) Door-open scenarios: Simulate changes in external conditions, such as sudden changes of room temperature resulting from unexpected door openings, aimed at evaluating the performance of the MPC controller under extreme circumstances.

By following this simulation plan, the performance of VAR-PSO-MPC can be evaluated in a CFD-simulated server room environment, and different scenarios can be tested for comparison with the baseline condition. The simulation will evaluate how the VAR-PSO-MPC framework performs in different contexts in terms of stability, accuracy, and efficiency. The results of these simulations will provide insights into the strengths and limitations of the framework and suggest areas for further improvement.

6.2.4 Simulation Tools and Techniques

1) Geometry tool: SOLIDWORKS. Build up a simulated 3D geometry environment of the server room, including all the physical space of the server room and the layout of the facilities.

2) Mesh tool: DistMesh package in STARCCM+: This is a popular open-source meshing package that provides functions for generating 2D and 3D Delaunay meshes. It includes several options for controlling mesh size and quality, as well as tools for visualizing and refining the mesh.

3) Define the physics in the server room: Computational Fluid Dynamics Toolbox in STARCCM+. This toolbox is specifically designed for simulating fluid dynamics and heat transfer in complex geometries. It includes a range of solvers for different types of flow problems, such as laminar and turbulent flow, and allows for the creation of custom models and user-defined functions.

4) CFD Simulation and MPC model verification: R- STARCCM+. This involves creating a simulation loop that updates the input variables and parameters at each time step, runs the MPC controller to determine the optimal cooling setpoints, and updates the simulation based on the controller output.

5) Evaluate the performance of the VAR-PSO-MPC controller in reducing the PUE of the simulated data centre environment. This may involve analysing key performance indicators such as the PUE, temperature gradients, and cooling system energy consumption, and comparing the

results with those obtained using the baseline setpoints.

6.3 Simulation modelling procedures

6.3.1 Geometry

In CFD simulations, one of the most crucial steps is to create a suitable geometry that accurately represents the physical domain. This process involves importing, evaluating, and simplifying complex geometries, and generating a suitable mesh.

The initial step involves the generation or acquisition of the geometric model, which can be achieved through the use of various tools such as Computer-Aided Design (CAD) software or other 3D modelling software. The geometric representation should accurately reflect the physical domain, incorporating pertinent details such as the object(s) dimensions and shape, as well as the surrounding environment. Subsequently, the geometry must undergo a thorough evaluation to address inconsistencies and guarantee its watertight integrity. Additionally, simplification procedures may be applied as necessary to reduce computational expenses.

In this experiment, the SOLIDWORKS software was employed to model the DC environment. To ensure accuracy and facilitate comparison with the prior experiment conducted in a physical DC located in Thailand, the server size and rack layout in the simulation were maintained at identical values. The historical data collected from the previous experiment was utilized to authenticate the current simulation environment. Additionally, the performance of the control framework was examined using various simulated scenarios that cannot be feasibly tested in a physical server room.

Figure 6-2 and Figure 6-3 demonstrate the geometry used for the simulation, with a top view and a 3D view respectively.

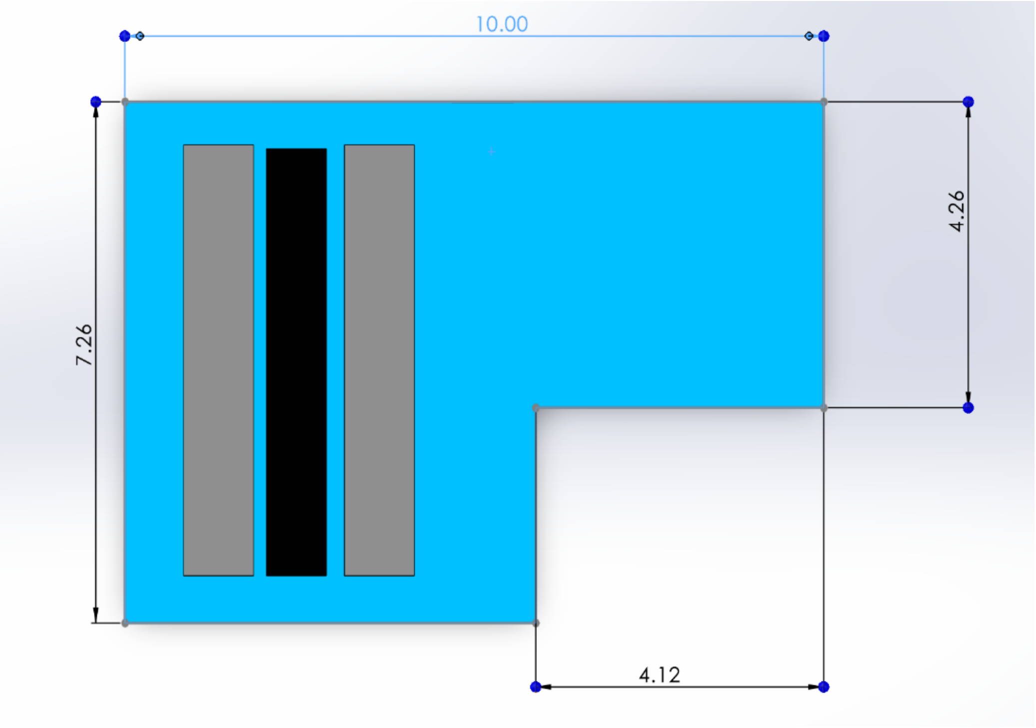


Figure 6-2 Top view of the server room geometry

6.3.2 Mesh generation and evaluation

7.2.2.1 Mesh generation

The fundamental objective of mesh generation is to partition the geometry of the problem domain into small, discrete elements to enable the numerical solution of the governing equations for fluid flow. It is imperative to note that the quality of the mesh plays a crucial role in determining the accuracy and reliability of the simulation results. Therefore, a well-designed mesh should strike a balance between high resolution in regions of interest and low resolution in regions of lesser importance. This not only reduces computational resources and time but also ensures the continuity of fluid flow across the computational domain.

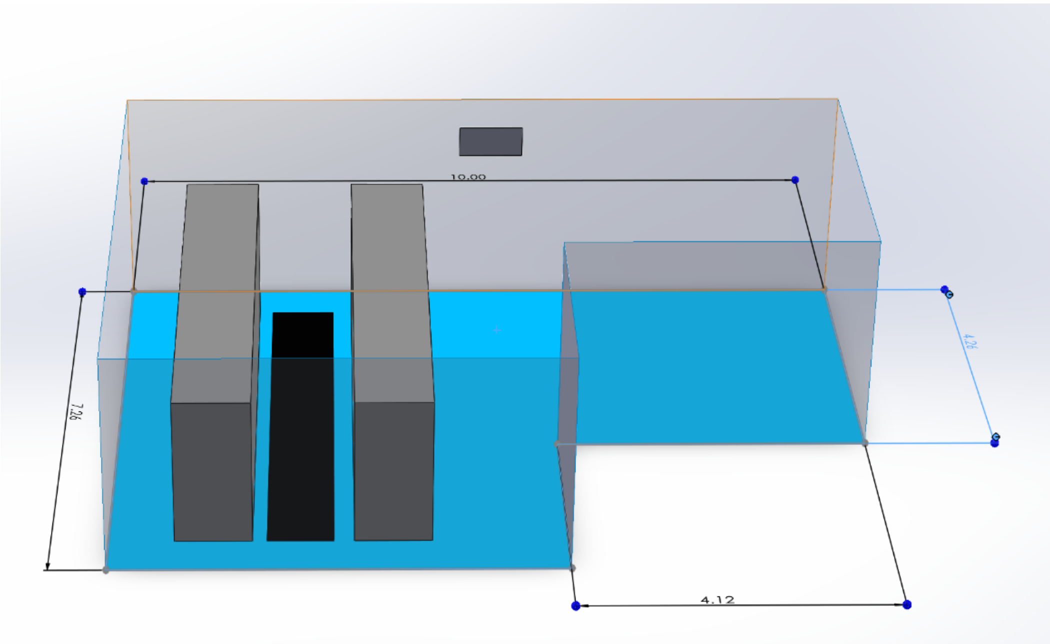


Figure 6-3 3D Server room geometry layout

The mesh type that has been chosen to simulate the DC environment is hexahedron (or "hex") mesh. The use of hex mesh in computational fluid dynamics (CFD) simulations of DCs offers several advantages. One primary advantage is its ability to accurately represent the complex geometry of the

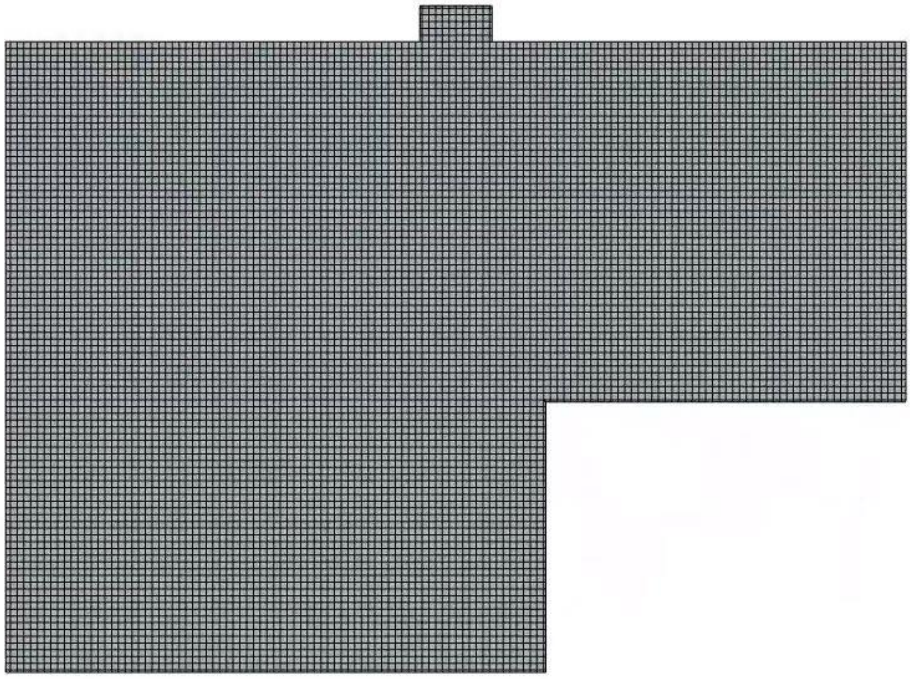


Figure 6-4 Top view of the server room mesh

components within the DC. This is due to the regular shape of hexahedral elements, which are cube-shaped with straight edges and right angles, providing a more uniform distribution of nodes and a more precise representation of the geometry. This contrasts with other mesh types, such as tetrahedral or prismatic meshes. Additionally, hex mesh has the potential to minimise numerical errors and provide better accuracy in regions of interest, such as where cooling systems or heat-generating components are located. This can lead to more reliable and accurate simulation results, aiding in the identification of potential thermal issues and enabling the design of effective cooling strategies. (Mittal & Iaccarino, 2005). Figure 6-4 shows the top view and Figure 6-5 shows the 3D view of the mesh.

The quality of the mesh plays a significant role in the accuracy of CFD simulations. A sensitivity test has been conducted to optimise the elements of the mesh.

7.2.2.2 Mesh sensitivity analysis

Table 6-1 presents the mesh quality in terms of a numerical index ranging from 0.000001 to 1, with higher values indicating better quality and lower values indicating poorer quality. The volume change range in a mesh is a metric used to evaluate the quality of the mesh. It is the ratio of the maximum volume change to the average volume change for all elements in the mesh. This range is calculated as the ratio of the maximum volume change to the average volume change across all elements in the mesh. A high-volume change range

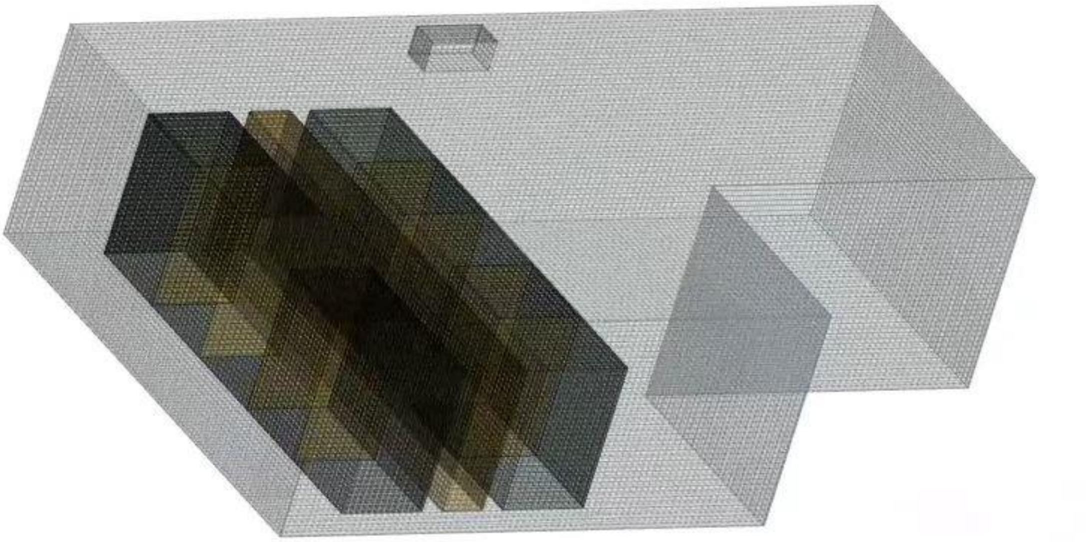


Figure 6-5 3D view of the server room mesh

suggests that certain elements in the mesh may be distorted, while a low volume change range indicates a more uniform shape across all elements.

Table 6-1 Numerical Indices of mesh quality

Volume change range	**Number of elements**	**Quality Index**
0.000000e+00<=Volumn change<1.000000-06	0	0.000%
0.000000e+00<=Volumn change<1.000000-05	0	0.000%
0.000000e+00<=Volumn change<1.000000-04	0	0.000%
0.000000e+00<=Volumn change<1.000000-03	0	0.000%
0.000000e+00<=Volumn change<1.000000-02	0	0.000%
0.000000e+00<=Volumn change<1.000000-01	5	0.001%
0.000000e+00<=Volumn change<1.000000-00	437080	99.999%

Additionally, the result of sensitivity analysis yielded an optimal mesh selection with a face value of 1270434 and vertices of 577568, which refer to the 2D polygons that constitute the surface of the 3D object being meshed. The faces and vertices together define the geometry of the mesh. Furthermore, we evaluated the validity of the mesh, and the results indicated that the mesh is topologically valid, with no negative volume cells. This evaluation process ensured that the mesh was of high quality and suitable for use in the CFD simulations.

6.3.3 Physical formulation of fluid properties

Fluid simulation for DC airflow by the CFD method involves solving the Navier-Stokes equations numerically to simulate the flow of air in and around the DC.

The formulations utilized for this simulation are as the following:

(1) Navier-Stokes equation

One common formulation used in CFD for simulating the airflow in DC is the incompressible Navier-Stokes equation, this equation describes the conservation of momentum for a fluid in terms of velocity, pressure, and temperature. In the context of CFD the Navier-Stokes equation is typically expressed in its dimensionless form, which is more suitable for numerical simulations.

$$\frac{\partial u}{\partial t} + u\nabla u = -\nabla p + v\nabla^2 u + f \tag{6.1}$$

Where:

u is the velocity vector of the fluid,

p is the pressure,

v is the kinematic viscosity,

f is any external forces acting on the fluid, and

∇ is the nabla operator.

The Navier-Stokes equation is supplemented by the continuity equation, which expresses the conservation of mass for the fluid.

(2) Continuity equation

The continuity equation is a fundamental equation in fluid dynamics that expresses the conservation of mass for a fluid. It states that the rate at which mass entries a control volume must be equal to the rate at which mass leaves the same volume, accounting for any accumulation or depletion of mass within the volume itself.

$$\nabla . \mu = o \tag{6.2}$$

(3) Energy function

The energy function (also known as the potential function), is a scalar field that describes the energy per unit volume of a fluid. The general form of energy function can be derived from the Navier-Stokes equation and continuity equation.

$$pc_p \frac{\partial T}{\partial t} + pc_p (u . \nabla) T = k \nabla^2 T + +Q + Q_{in} - Q_{out} \tag{6.3}$$

Where:

T is the temperature,

$pc_p$is the specific heat capacity,

k is the thermal conductivity,

Q is any internal heat sources or sinks,

Q_{in} is the amount of heat flowing into the system,

Q_{out} represent the heat flowing out of the system.

For the CFD simulation of the server room, we will modify the above equation in terms of AC and server CPU usage.

$$pc_p\left(\frac{\partial T}{\partial t}+u.\nabla T\right)=k\nabla^2 T+Q_{server}+Q_{AC} \quad (6.4)$$

Where:

Q_{AC} is the rate of heat transfer per unit volume due to the air conditioners.

Q_{server} is the rate of heat transfer per unit volume due to the servers.

To incorporate the heat generated by CPUs, we could model each CPU as a point source of heat within the computational domain. The rate of heat transfer per unit volume is due to the CPUs. Q_{server} could then be expressed as:

$$Q_{server}=\sum_{i=1}^{N} q_i\,\delta(X-X_i) \quad (6.5)$$

Where :

N is the number of CPUs in the domain,

q_i is the rate of heat generation per unit volume for the i-th CPU,

X_i is the location of the i-th CPU,

$\delta(X-X_i)$ is the Dirac delta function, which is zero everywhere except at $X=$

X_i where it has an infinite value that integrates to 1. This means that there is no heat generation at any point in the domain except at the location $X = X_i$.

This modified energy equation takes into account the heat generated by CPUs and can be solved numerically along with the Navier-Stokes equations and mass conservation equation to simulate the fluid dynamics and temperature distribution in a DC server room.

(4) Turbulence equation

According to a study by Wu et al. (2019), the k-epsilon turbulence model is commonly used in CFD simulations of server rooms to accurately predict the distribution of temperature and air velocity within the room. The model solves transport equations for the turbulent kinetic energy and the rate of dissipation of turbulent kinetic energy, which takes into account the production and dissipation of turbulence due to fluid flow and shear stresses, as well as the diffusion of turbulence through the fluid. Empirical constants are used to account for the effects of turbulence on the fluid, and these can be adjusted to match experimental data or to better match the specific conditions of the server room being simulated. Overall, the k-epsilon turbulence model is a useful tool for simulating the complex fluid dynamics of DC server rooms. Based on the assumption that the behaviour of turbulence at a given point in space is influenced by the conditions at nearby points, the turbulence equations adopted for this simulation can be denoted as the following:

$$\frac{\partial \rho k}{\partial t}+\frac{\partial \rho k u_i}{\partial x_i}=\frac{\partial}{\partial_{xj}}\left[\left(\mu+\frac{u_t}{\sigma k}\right)\frac{\partial k}{\partial x_j}\right]+Pk-\rho\varepsilon \tag{6.6}$$

$$\frac{\partial \varepsilon}{\partial t}+\frac{\partial \rho \varepsilon u_i}{\partial x_i}=\frac{\partial}{\partial_{xj}}\left[\left(\mu+\frac{u_t}{\sigma \varepsilon}\right)\frac{\partial k}{\partial x_j}\right]+C_{l\varepsilon}\frac{\varepsilon}{k}Pk-C_{2\varepsilon}\rho\frac{\varepsilon^2}{k} \tag{6.7}$$

Where:

k is the turbulent kinetic energy,

ε is the rate of dissipation of turbulent kinetic energy,

ρ is the density of air,

u_i is the velocity vector,

μ is the molecular viscosity of air,

μ_t is the turbulent viscosity,

σk and $\sigma \varepsilon$ are empirical constants,

Pk is the rate of production of turbulent kinetic energy,

$C_{l\varepsilon}$ and $C_{2\varepsilon}$ are additional empirical constants that help to account for the effects of turbulence on the fluid,

x_i and x_j denotes the spatial location in the domain where the turbulent flow is being simulated.

The above equations provide the mathematical framework of the air and heat transfer in the server room, which can be used to formulate the CFD simulation. The energy equation describes how heat is generated by the CPUs and removed by the air conditioning system, while the turbulence model accounts for the complex, three-dimensional flow patterns that arise in the presence of obstacles and thermal gradients. Moreover, the effectiveness of the k-epsilon model in simulating flows in enclosed spaces is widely recognised, making it a commonly used turbulence model in the field of computational fluid dynamics. By numerically solving these equations using a suitable CFD solver, valuable insights can be obtained into the thermal behaviour of the server room. These insights can then be used to optimise the design of the cooling system, ensuring the reliable operation of the servers while minimising energy consumption. Such an approach holds promise for improving the overall efficiency and sustainability of DC operations.

6.3.4 Server room attributes and boundary conditions

To ensure the reliability and performance efficiency of the IT infrastructure in the simulation, the room size, the number of server racks, total power capacity, air conditioner capacity, server thermal load, and power usage effectiveness (PUE) are among the key attributes that must be carefully evaluated during the planning and design phases. Additionally, boundary conditions such as the inlet air temperature and airflow rate must be closely monitored and controlled to maintain optimal conditions for the servers and other IT equipment. In the following section, we will provide a detailed discussion of the essential attributes and boundary conditions for the successful operation of a server room, based on the latest thermal guidelines and industry best practices.

7.2.2.3 Server room attributes

Table 6-2 presents the essential attributes of the server room that must be considered when conducting a simulation. The size of the server room used in the simulation is identical to that of the actual experiment conducted earlier, measuring 57.43 square metres with a height of 3 metres. The table includes information on the number of server racks, total power capacity, air conditioner capacity, cooling capacity, and server thermal capacity. These attributes are crucial in accurately modelling the temperature and airflow distribution in the server room, as they affect the thermal load and cooling requirements of the IT equipment. Additionally, power usage effectiveness (PUE) is also listed as a key attribute, as it reflects the overall efficiency of the power and cooling systems in the server room. Other relevant attributes, such as the location of the server room, wall and floor construction, and airflow obstructions, may also impact the simulation results and should be carefully considered. Boundary conditions will be calculated based on the key attributes that are given.

Table 6-2 Key attributes of the server room

Attribute	**Value**
Room size	57.m2
Room height	3 m

Number of server racks	10
Air conditioner capacity	2 x 14,000 W
Server thermal capacity	10 x 5,200 W
General Power usage effectiveness (PUE)	2.3

7.2.2.4 Boundary conditions

Assumptions

A. *The air conditioners provide all of the cooling for the server room.*

B. *Incompressible flow: The simulation assumes that the air inside the server room is incompressible, meaning that changes in density due to changes in temperature or pressure are negligible.*

C. *No radiation effects: The simulation assumes that the effects of radiation on the heat transfer within the server room are negligible.*

(1) Total thermal load

$$Total\ thermal\ load$$
$$= server\ thermal\ capacity$$
$$+ air\ conditioner\ cooling\ capacity$$

Under the general server room condition,

$$Total\ thermal\ load = 52000W + 28000W$$

$$Total\ thermal\ load = 80000W$$

(2) Mass flow rate

Effective cooling of a server room is heavily dependent on the rate of airflow, making it a critical factor in maintaining optimal temperatures for IT equipment. The airflow rate can be calculated using the following formula:

$$Airflow\ rate\ (m^3/h)$$
$$= Total\ thermal\ load\ (W)\ /\ (1.2\ \times\ Specific\ Heat$$
$$\times\ Temperature\ Difference)$$

Assuming a specific heat of $1.0\ kJ/kg * K$ and a temperature difference of $0.88°C$, which means that for every kilogram of air passing through the system, there is a change in enthalpy (heat content) of approximately $10.05\ kJ$, and the difference between the inlet and outlet temperatures of the air passing through the system is $3°C$, the required airflow rate is:

$$Airflow\ rate\ =\ 80000W\ /\ (1.2 \times\ 1.0kJ/kg \times K\ \times\ 3°C$$
$$=\ 22222\ m^3/h$$

So the mass flow rate of air required to achieve the required airflow rate would be:

$$Mass\ flow\ rate\ (kg/s)$$
$$=\ \frac{Airflow\ rate\ (\ m^3/h)}{3600} \times\ Air\ Density\ (kg/m^3)$$

Assuming an air density of $1.2\ kg/m^3$, the required mass flow rate is:

$$Mass\ flow\ rate\ =\ \frac{22222m^3/h}{3600} \times\ 1.2\frac{kg}{m^3} =\ 7.4kg/s$$

To achieve this mass flow rate, the airflow should be divided equally between the two ACs, so each AC needs to provide:

$$Mass\ flow\ rate\ per\ air\ conditioner\ =\frac{7.4\ kg/s}{2} = 4.7kg/s$$

(3) Inlet air boundary condition

Following the guidelines by ASHRAE (2021) for DC cooling, the recommended range for an inlet air temperature of the air conditioning units is between 18°C and 27°C. The calculations below are to determine the required cooling capacity of each AC unit to achieve the desired inlet air temperature and maintain the appropriate temperature difference between the inlet air and room temperature. This information is important for selecting appropriate air conditioner units and designing an effective cooling system for the server room. Assuming an inlet air temperature of 22°C and a room temperature of 25°C, the air conditioners need to cool the air by:

$$\begin{aligned} &Temperature\ difference \\ &\qquad = Inlet\ air\ temperature\ -\ Room\ temperature \\ &\qquad = 22°C\ -\ 25°C = -3°C \end{aligned}$$

Therefore, each air conditioner needs to provide a cooling capacity of:

$$Cooling\ capacity\ per\ air\ conditioner$$
$$= \ Mass\ flow\ rate\ per\ air\ conditioner \times \ Specific\ heat$$
$$\times \ Temperature\ difference$$
$$= 7.4 kg/s \ \times \ 1.0 \ kJ/kg * K \ \times (-3°C) = \ -22.2 \ kW$$

Since the cooling capacity cannot be negative, therefore here should assume that each air conditioner needs to provide a cooling capacity of 22.2 kW.

With the distance between the AC units and the inlet air point of the server room, airflow velocity, and ACs cooling capacities, the inlet air temperatures can be calculated by the following equations:

$$Q \ = \ m \ \times \ c\_p \ \times \ \Delta T \qquad (6.8)$$

Where:

Q = Cooling capacity of the air conditioning unit (kW)

m = Mass flow rate of air (kg/s)

c_p = Specific heat of air

$(kJ/kg \times K)$ ΔT = Temperature difference between inlet and outlet air (K)

Assuming that the cooling capacity of the air conditioning unit, the airflow velocity, and the distance between the AC units and the inlet air point, the mass flow rate is determined, the inlet air temperature can be calculated using the following formula:

$$T\_inlet = T\_outlet - (Q / (m \times c\_p)) \quad (6.9)$$

Where:

$T\_inlet = Inlet\ air\ temperature\ (°C)$

$T\_outlet = Outlet\ air\ temperature\ (°C)$

(1) Fan speed dynamic condition

To sustain the necessary mass flow rate, the fan speed of the ACs must be regulated according to their performance curves. These curves can be obtained through the manufacturer's specifications or by conducting experiments. In addition, to ensure the required mass flow rate and inlet air temperature are maintained, the fan speed must be dynamically adjusted based on the thermal load of the server room.

In summary, the inlet boundary conditions for the server room are:

(2) Humidity

It is generally recommended to maintain the humidity within the range of 40% to 60% to avoid equipment damage from condensation or static electricity build-up (ASHRAE,2021). Maintaining proper humidity levels helps to prevent equipment failures and data loss. For this simulation, a constant operating humidity of 50% was chosen to simplify the model.

(3) AC Airflow velocity

ASHRAE (2021) has outlined certain recommendations regarding air velocity in DCs, suggesting that the optimal range of air velocity is typically between 0.15 and 0.76 meters per second (30 to 150 feet per minute) for a majority of cooling equipment in a server room. Typically, the calculation of AC airflow velocity given the mass flow rate would be presented as the following:

$$velocity = mass\ flow\ rate / (density \times cross\_sectional\ area)$$

For this simulation, a constant air velocity of 0.25 meters per second was utilized as a simplifying assumption.

(4) Wall boundaries

The wall surface is assumed to have a thermal conductivity of 1.5 W/mK.

The thickness of the wall is assumed to be 0.2 metres.

The air outlet point is set to be the freedom outlet (pressure outlet with 0 pressure.

These boundary conditions can be used to simulate the behaviour of airflow and temperature around the air outlet point on the wall of the server room.

(5) The Porous surface boundary

The AC airway located underneath the raised floor is designed to distribute the cool air evenly across the room. In this CFD simulation, the airway is modelled as a porous surface with a porosity of 90%. This means that 90% of the surface area is covered by small openings or pores that allow the air to flow through, while the remaining 10% is assumed to be impermeable. The porous surface is modelled using a Darcy-Brinkman model which accounts for the pressure drop and flow resistance due to the presence of the porous surface. The porous boundary condition of the airway allows for accurate modelling of the airflow and temperature distribution in the DC, ensuring efficient cooling of the servers.

6.3.5 The CFD solver

(5) Solver Selection and Configuration

The transient solver, specifically tailored to capture the unsteady characteristics of the flow and temperature fields in a server room, has gained significant adoption in the simulation of dynamic server room states. To ensure the attainment of simulation results that meet the required levels of accuracy, it is recommended that the transient solver settings be appropriately configured following the size and complexity of the server room geometry. The versatility of the continuum mechanics solver in STAR-CCM+ enables it to solve a broad spectrum of fluid dynamics problems, encompassing laminar and turbulent flows, multiphase flows, heat transfer, and combustion. Thus, it has been selected to tackle the algorithms outlined earlier. The solver utilizes a cell-centred approach and solves the equations of fluid motion, including the Navier-Stokes equations, for each cell in the discretized domain. The solver includes advanced numerical algorithms such as the pressure-implicit with the splitting of operators (PISO) algorithm, which is used to solve the pressure-velocity coupling in unsteady flows, and the Menter-SST (Shear Stress Transport) turbulence model, which is widely used for modelling turbulent flows in complex geometries.

(6) Time Step and Maximum Number of Iterations

The time step governs the pace of the simulation and has a direct bearing on the accuracy and stability of the obtained results. To ensure the attainment of simulation results that are commensurate with the desired level of accuracy, the selection of an appropriate time step size warrants careful consideration of the physical attributes of the server room. Typically, a smaller time step size would yield more precise results, however, albeit at the cost of heightened computational time. The maximum number of iterations is the maximum number of time steps that the solver will perform during the simulation. The appropriate setting of a value is of utmost significance in guaranteeing the timely completion of a simulation, whilst concurrently capturing the transient behaviour of flow and temperature fields. The maximum number of iterations should be determined based on the server room's size and complexity, as well as the desired level of precision.

Furthermore, a fixed time step provides the simulation process predictability and stability in contrast to the variable time step. It ensures uniformity in the time intervals at which calculations are performed, therefore simplifying the computational procedures and enhancing the overall stability of the simulation. This is particularly crucial in systems with complicated dynamics that are susceptible to environmental changes, such server rooms.

In this simulation, a time step of one second has been selected, along with a maximum iteration limit of 100, following comparative testing of various time intervals and iterations. This decision was made to balance system runtime and performance considerations.

(7) Convergence Criterion

The convergence criterion is a measure of the accuracy of the simulation results. It is the threshold value that the solver uses to determine when the solution has reached a steady state. A suitable convergence criterion for a dynamic state server room simulation is a maximum residual error of 0.01 or less. This means that the solution has reached a steady state when the maximum residual error in the solution is less than 0.01. However, the convergence criterion should be chosen based on the desired level of accuracy and the size and complexity of the server room geometry.

6.4 MPC performances under different simulated scenarios

To assess the efficacy of the VAR-PSO-MPC control framework in minimising the Power Usage Effectiveness (PUE) of a server room, a series of simulations were conducted under different scenarios. The objective of these simulations was to validate the performance of the control framework by analysing the impact of different parameters on the PUE, including server utilisation, cooling system efficiency, and ambient temperature. The simulations were performed using a detailed model of the server room, which included the thermal and energy characteristics of the equipment, as well as the cooling infrastructure. The results of the simulations were analysed to determine the effectiveness of the VAR-PSO-MPC control framework in reducing PUE under different operating conditions. The details of the simulations and the corresponding results are presented in the subsequent sections.

6.4.1 Baseline condition

This section presents the baseline condition for the server room simulation, which is based on the geometry and mesh that were generated in the previous section. The purpose of this baseline condition is to establish a reference point for future simulations and to provide a benchmark for evaluating the effectiveness of the control strategies under various conditions.

The baseline condition assumes that the server room has been operated under normal conditions without the manipulation of the server workload and ACs temperature settings for 12 hours during the daytime. The initial configuration of the server room is based on the results of a previous experiment from the real-world DC. It is assumed that the cooling system operates with 100% efficiency, and the initial supply air temperature for both ACs is 17°C. The IT load has been allocated to roughly 20% of the total capacity, which is 10000 watts. It is commonly observed in the literature that DCs usually function at less than 30% of their total capacity. Notably, the calculated air outlet temperature of the air conditioning (AC) system deviates from the AC setpoint temperature by 0.88°C, which can be attributed to the spatial separation (2 metres) between the two. In addition, the initial temperature and humidity levels within the server room were also set to match observed conditions, and the power consumption of the IT equipment and cooling system was based on actual usage data. Moreover, the initial placement and configuration of the IT equipment and cooling infrastructure were modelled after the real-world DC layout, and the initial conditions of the external environment, such as outdoor temperature and humidity levels, were taken into consideration. the server room conditions in the simulated environment adhered to the These settings were chosen to ensure that the simulated environment closely resembled real-world DC conditions and that the results of the study were relevant to practical DC operations.

Figure 6-6 Temperature distribution and fluid contour under the baseline conditiondisplays a graphic visualisation of the fluid flow and temperature contour at the baseline condition. The figure provides a clear representation of the fluid's motion and highlights important features, such as velocity, vortices and boundary layers. Areas of high velocity typically exhibit rapid and smooth streamlines, indicating the presence of high fluid momentum. Conversely, areas of low velocity may display turbulent or irregular flow patterns and slower streamlining. Moreover, vortices will show as rotating structures within the fluid flow and are characterised by their circular or spiral flow patterns. Moreover, variations in temperature between servers indicate the variation in heat generation due to the difference between the workload that has been assigned. Additionally, high PUE constantly around 2.3 infers that under normal operating conditions, the room has been over-cooled, and energy waste exists.

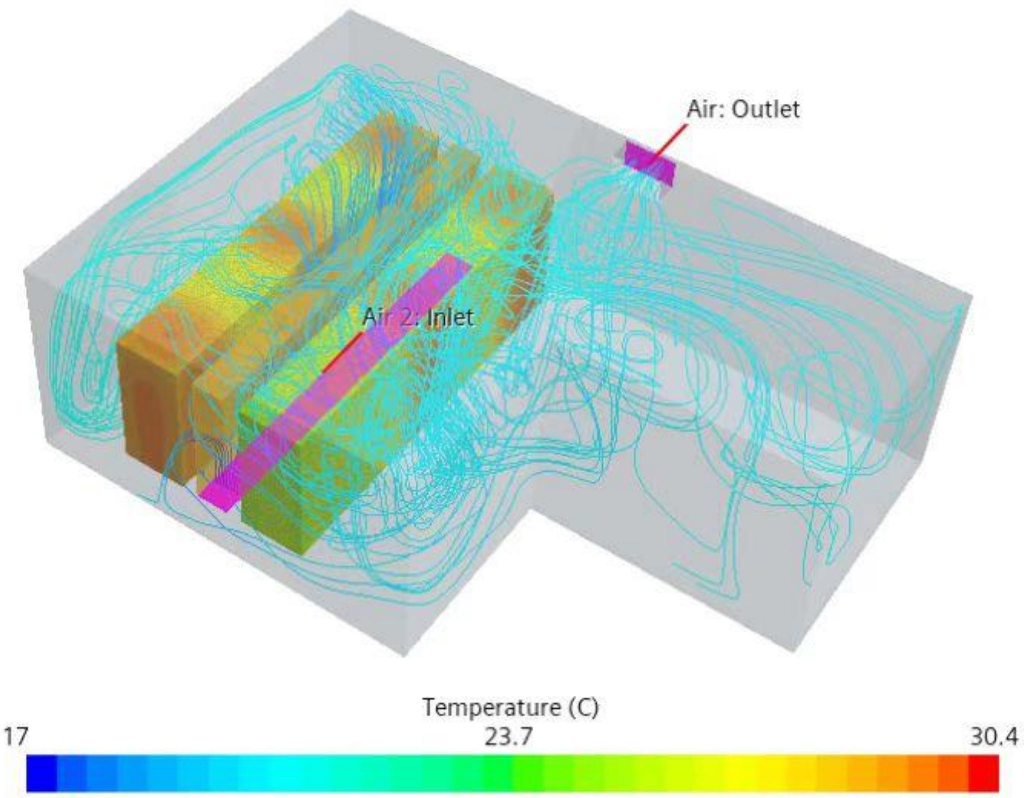


Figure 6-6 Temperature distribution and fluid contour under the baseline

Figure 6-7 and Figure 6-8 demonstrates the servers working temperature and the ambient temperature under baseline conditions, where the DCs typically operate at a constant temperature without temperature control. The initial temperature settings of both ACs were 17°C, which led to an over-cooling problem in the server room. Despite the significant workload assigned during busy working hours, the temperature in the server room remained notably lower than the upper boundaries, which are 27°C for the ambient temperature 32°C for the servers' temperature, leading to inefficient energy usage. This highlights the need for improved strategies to optimise the energy efficiency of server rooms, even under high workloads.

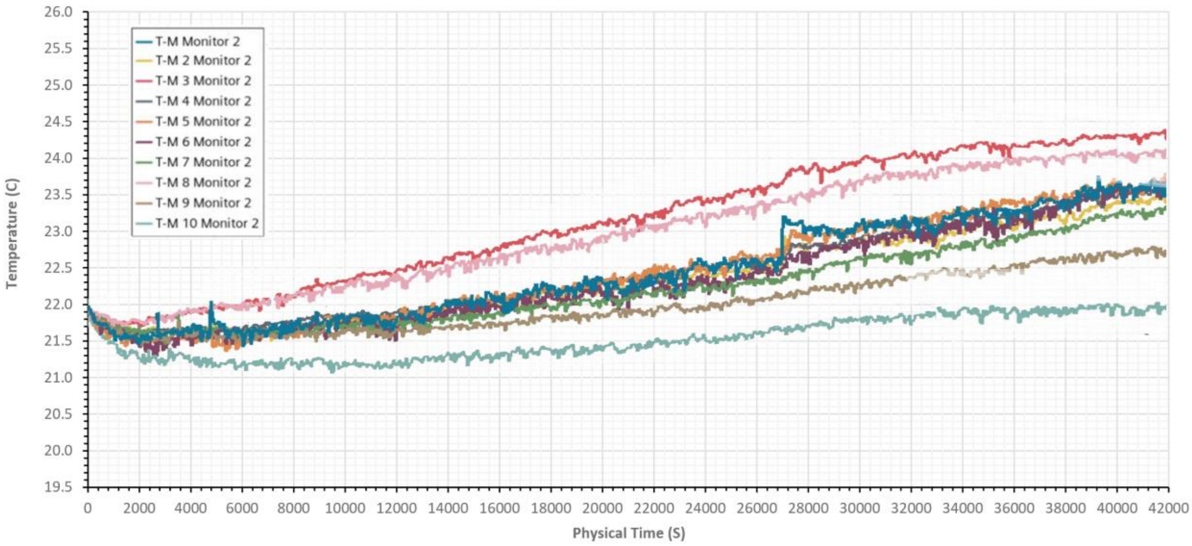


Figure 6-7 The servers' temperature under baseline condition

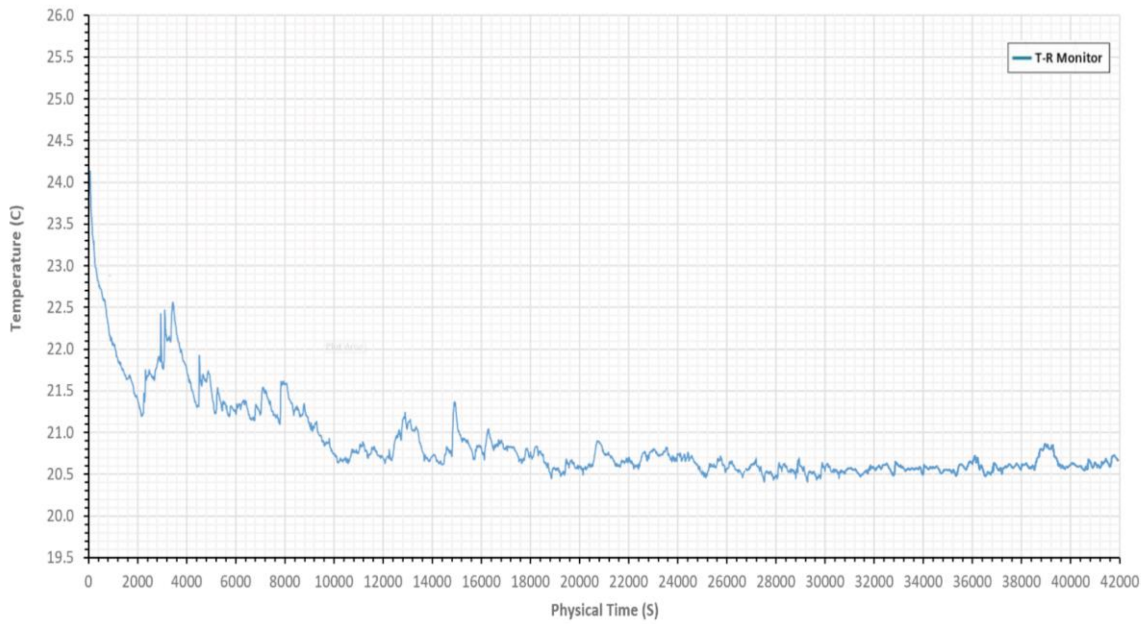


Figure 6-8 The ambient temperature under baseline condition

6.4.2 Server load variation scenario

This simulation scenario includes variations in server workload to reflect real-world operating conditions. A series of simulations will be conducted to assess the system's performance under different workloads. The CPU workload will be varied from 19% to 70% of the total capacity in increments of 10% at each change. The simulation will be executed for 12 hours to capture the transient behaviour of the system. Figure 6-9 depicts the temperature distribution and fluid contour under workload variation scenario by simulation.

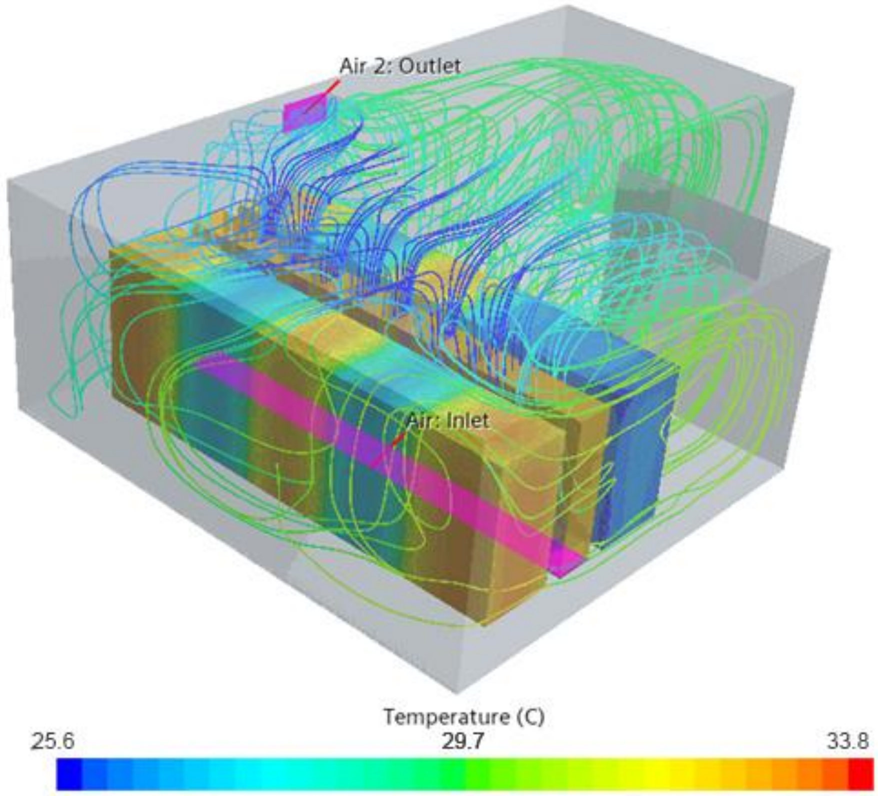


Figure 6-9 Temperature distribution and fluid contour under servers workload variation scenario

The presentation of the simulation results will involve an evaluation of DC variables, involves temperature distribution, server temperature, ambient temperature, and PUE, in comparison to the corresponding baseline simulation outcomes. The server temperature changes over 12 hours. Figure 6-10 demonstrate the effects of server workload variation. To maintain optimal conditions, the proposed MPC control strategy has been employed to execute the optimal ACs temperature setpoints and regulate the ambient and server temperature levels within a specific range, while simultaneously minimising

PUE, it also demonstrate a scenario where the temperature of the servers is close to exceeding the temperature limit of the servers (32°C) from the beginning of the 11th hour of simulation, indicating that the current maximum cooling capacity of the server room cannot handle the workload increase beyond 67%. Figure 6-11 shows the ambient temperature under the MPC control over time, provides evidence of the efficacy of MPC control in regulating ambient temperature during workload variations. The graph illustrates that, whenever the workload increased, the ambient temperature experienced severe surges. However, with the application of MPC control, the temperature was brought down to a static stage within approximately half an hour. However, with the workload reaching 67%, the temperature displays a trend towards approaching the upper limit of 27 degrees.

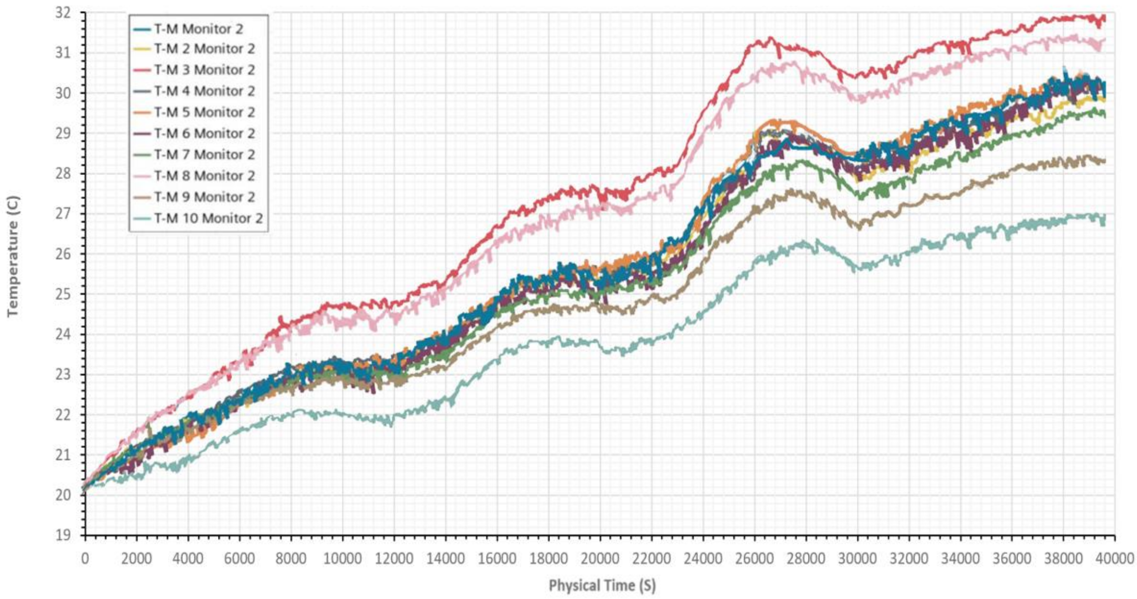


Figure 6-10 Servers' temperatures with workload variation

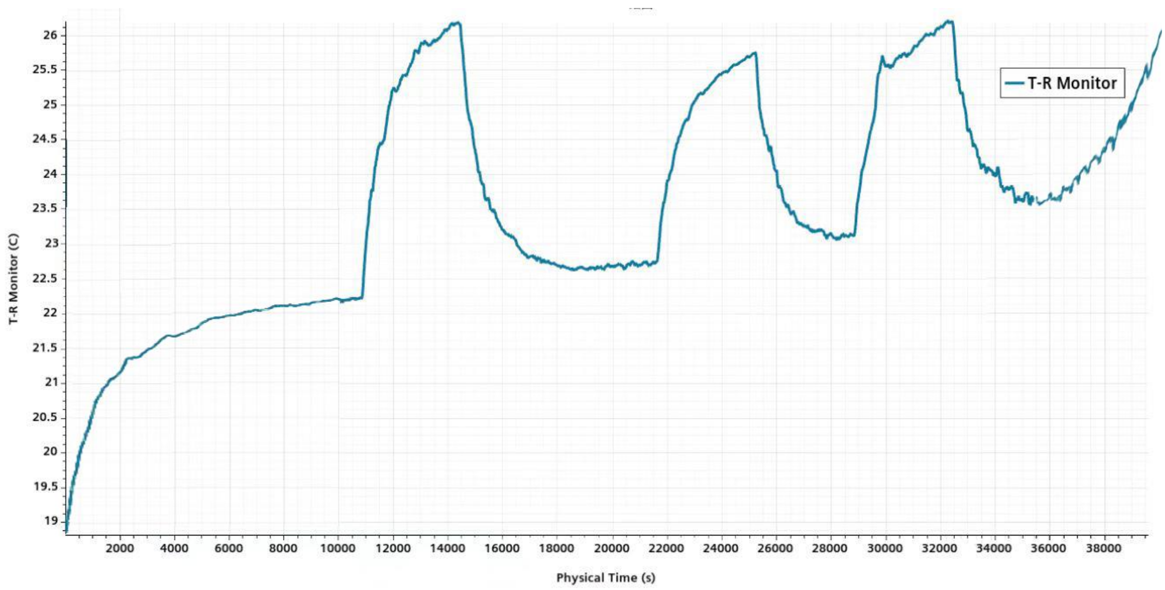


Figure 6-11 Ambient temperature with workload variation

Figure 6-12 displays the movements of the CPU, PUE, and temperature settings for the ACs. The results of the CFD simulation indicate that the VAR-PSO-MPC control strategy effectively adjusts the two ACs 'temperature settings based on predicted server temperatures, ambient temperatures, and PUE. The simulation involved gradually increasing the server workload every few hours from 19.23% to 67.31%, and the results demonstrate that the implementation of the MPC strategy yields commendable outcomes in reducing PUE and ensuring that all temperature measurements remain within acceptable limits. To investigate whether the reduction in PUE was primarily attributed to the control strategy and not the addition of IT workload, the CPU workload was held constant for the initial three-hour period spanning 10,800 seconds. Specifically, the implementation resulted in a significant decrease in PUE from its initial value of 2.30 to a final value of 1.9046 when the IT workload remain stable, indicating the efficiency of the MPC control strategy.

Nonetheless, when the IT workload reaches 67.31%, the heavily loaded servers were predicted to approach the upper-temperature threshold, despite efforts to optimise the ACs' temperature setpoints based on VAR predictions, wherein both ACs were set at a minimum temperature of 16°C. These results imply that the present air conditioning capabilities are insufficient to accommodate the temperature rise resulting from high processing loads on the servers. This suggests that at this stage, additional cooling devices or an increase in the cooling capacity of the current devices is necessary to ensure that the system remains within the safe operating temperature range.

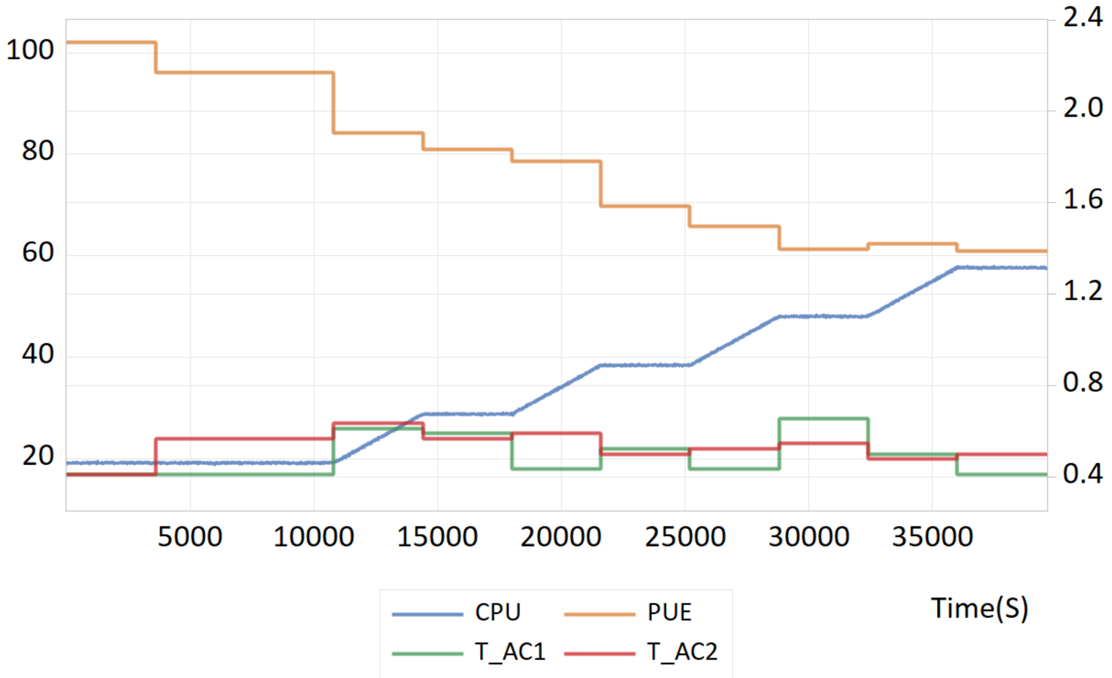


Figure 6-12 PUE and CPU with MPC control under workload variation

Figure 6-13 illustrate the servers' temperature for the last two hours under a workload variation scenario. The yellow zone in the figure represents a prediction that the server with the highest workload will reach the temperature upper boundary at 39651 seconds. In the actual simulation environment, the temperature upper boundary was reached at 39702 seconds, 51 seconds later than the predicted time. Despite this error, the advance notice provided by the prediction allowed the control system to respond to the temperature alarm promptly. To provide additional evidence supporting the prediction and to evaluate the behaviour of servers' temperature, the simulation has been conducted for a full period of 12 hours. The results indicate a critical consequence: Without a server down-locking system, when the workload assigned to the server exceeds its capacity and the cooling capacity of the air conditioning units is insufficient, the temperature of the server quickly becomes unstable and the over-heating of the servers may lead to potential performance issues or even system failure and hardware damage.

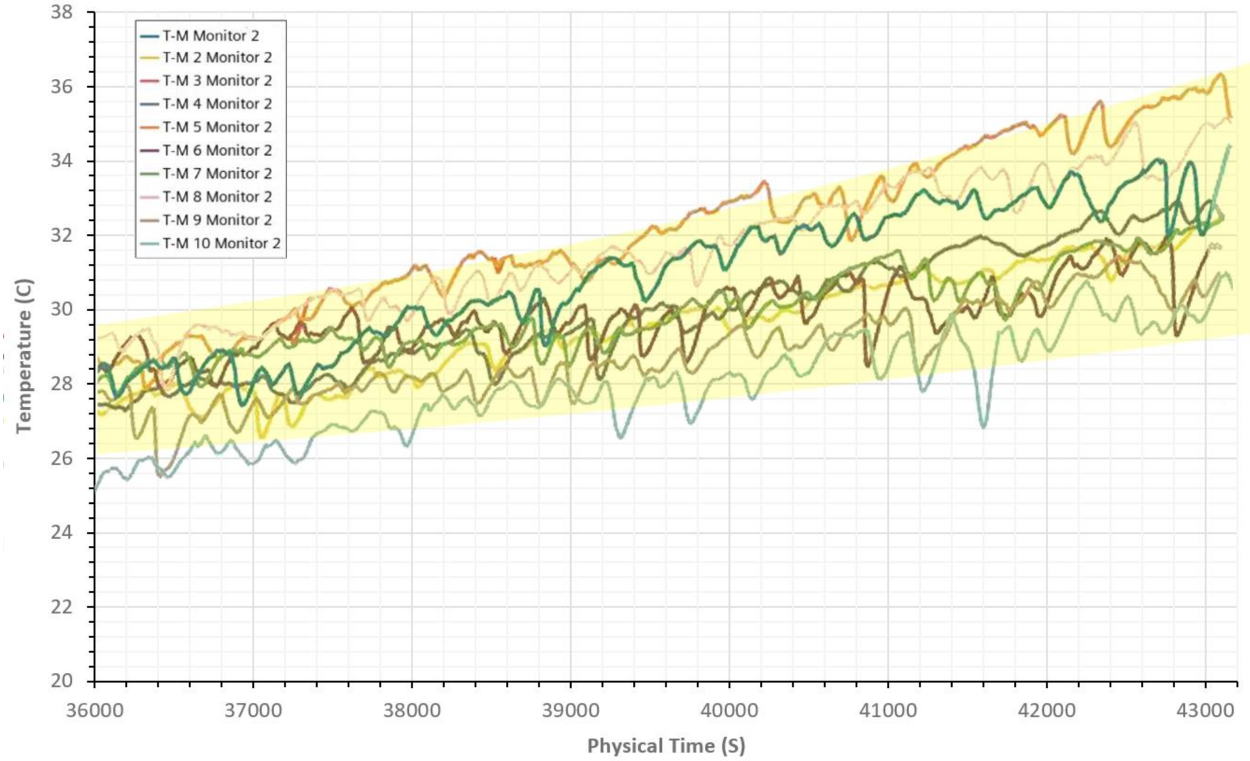


Figure 6-13 Temperature boundary alarm of the server temperatures

Further analysis of the simulation results revealed that the temperature distribution was not uniform, with some areas experiencing higher temperatures than others. This highlights the importance of accurate prediction and control of temperature distribution to ensure the optimal performance of the DC. The simulation also showed that the ambient temperature had a significant impact on the system's performance, indicating the need for a reliable and robust cooling system to handle fluctuations in the ambient temperature.

In conclusion, the results of the CFD simulation suggest that the VAR-PSO-MPC control strategy is effective in managing the PUE and maintaining the temperature within the boundary conditions under normal workload conditions. However, additional cooling capacity or cooling devices may be necessary to handle higher workload conditions. The simulation results also emphasise the importance of accurate temperature prediction and control and the need for a reliable cooling system to ensure optimal DC performance.

6.4.3 Scenario with additional cooling unit

In this section, the scenario of adding supplementary AC units to the existing infrastructure of the current DC scenario will be presented. The aim is to evaluate its impact and assess the system's adaptability to changes.

As evidenced by the previous scenario, which involved workload fluctuations, an insufficient-sized air conditioning system may result in overheating issues when a heavy workload is assigned to the server. Therefore, incorporating an extra air conditioning unit into the present infrastructure is imperative to ensure that the servers operate safely and within acceptable temperature limits.

Figure 6-14 depicts the comparison of the temperature of a heavy-load server that operated under two ACs and three ACs scenarios respectively. As the result demonstrated in the preceding section, the temperature of the heavy-loaded servers touched the upper boundary when the CPU load is up to 67.31% (shown as the red line). Alternatively, additional cooling power from the supplementary AC has been applied to cool down the servers as an emergency option. As the figure illustrated, combined with the MPC control the server temperature has been significantly reduced from 28°C to 26°C, followed by a gradual rise resulting from accumulated heat, which nevertheless remained below the upper-temperature threshold.

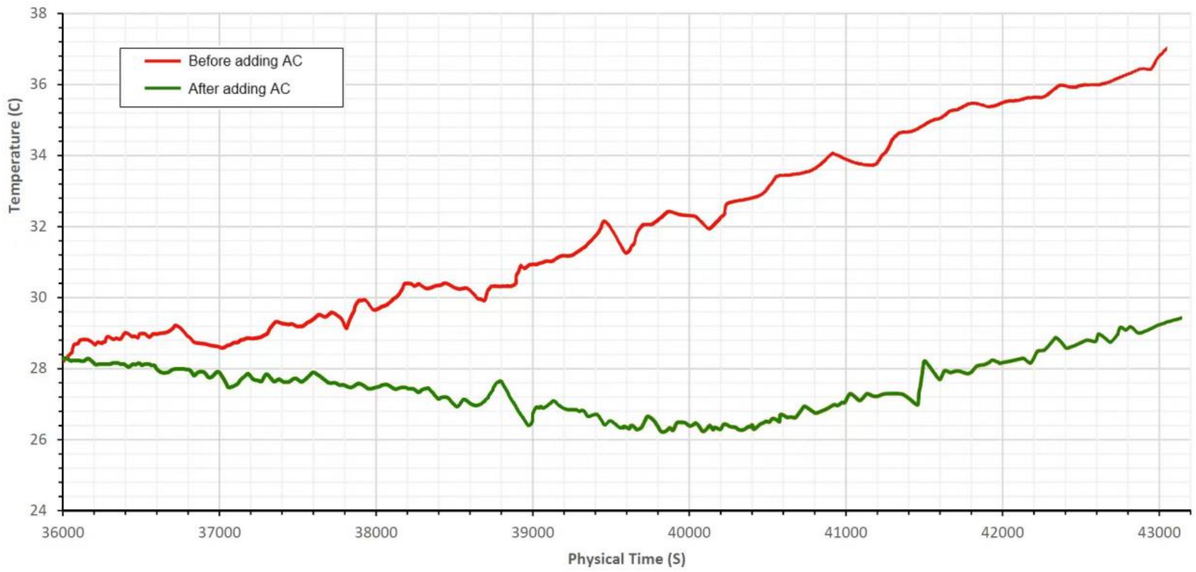


Figure 6-14 Temperature of the heavily loaded server with two and three supplementary Air Conditioning Units Applied

Figure 6-15 illustrates the PUE movements following the adjustment of ACs temperature settings. Given the VAR model's prediction that the temperature of the heavily loaded server would surpass the threshold at the 11th hour, mitigation measures to activate additional air conditioning must be implemented to curtail the temperature. The figure presents compelling evidence that the incorporation of supplementary air conditioning yielded a substantial increase in the PUE metric, rising from 1.4 to 2.2 in the initial hour of operation. However, through the utilisation of MPC control, the PUE value has been progressively reduced to 1.7 by the end hour of operation. Furthermore, the PUE readings remained stable from the ninth (physical time 32400s) to the twelfth (physical time 43200s) hour, indicating that the MPC control strategy is capable of reducing PUE by 0.5 within a span of 7 hours, particularly under critical server temperature conditions.

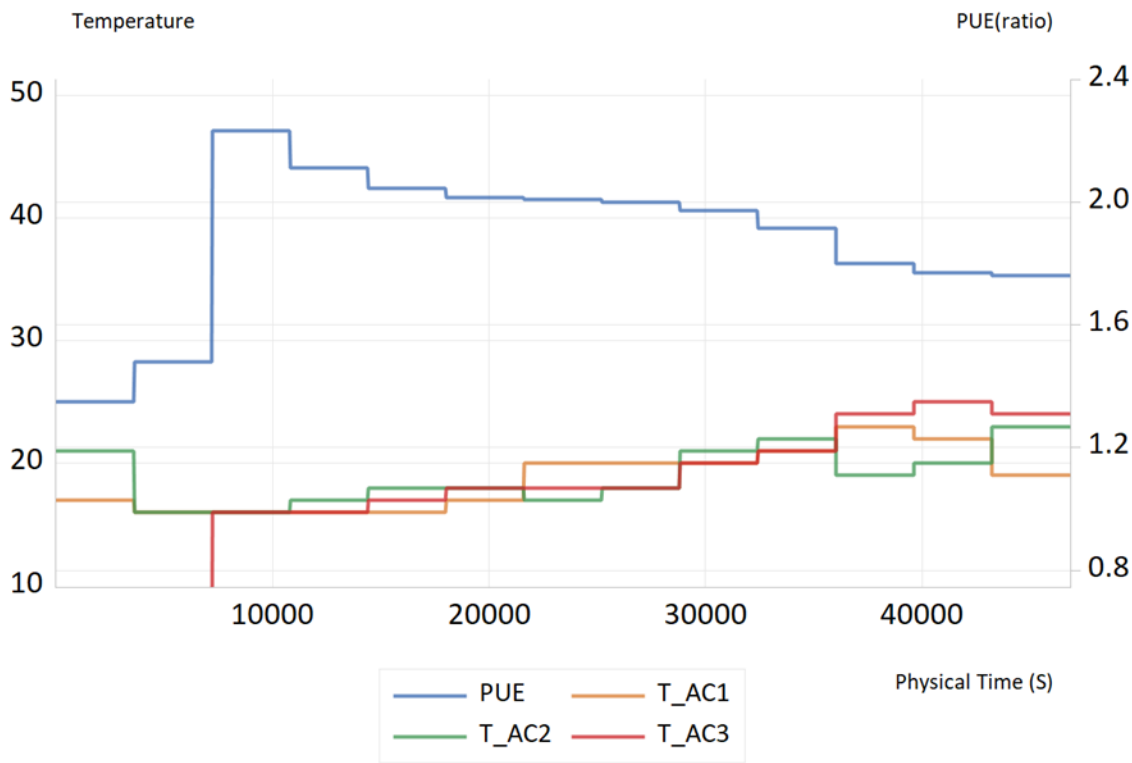


Figure 6-15 ACs setpoints and PUE movements with the additional AC

6.4.4 Scenario with additional server racks

Based on the modification of AC units in the previous section, the following section will introduce the scenario of adding servers to the existing infrastructure. To accommodate the additional servers, an AC airway will also be installed to facilitate the efficient inlet of air to the server room. This alteration aims to evaluate the capability of the proposed control strategy to adapt to changes in the infrastructure, specifically those that involve significant alterations to the geometry. Additionally, the incorporation of additional servers will function as a stress test for the control strategy, as it will be required to adapt to the amplified thermal loads and energy consumption. This highlights the significance of the capability if the control strategy in response to changes in infrastructure, emphasising its crucial role in managing the efficient operation of the system.

6.4.4.1 Geometry Mesh Design and Sensitivity Analysis

(1) Geometry changes and mesh designing

Figure 6-16 depicts the 3D view of the geometry layout of the server room with additional servers and airways.

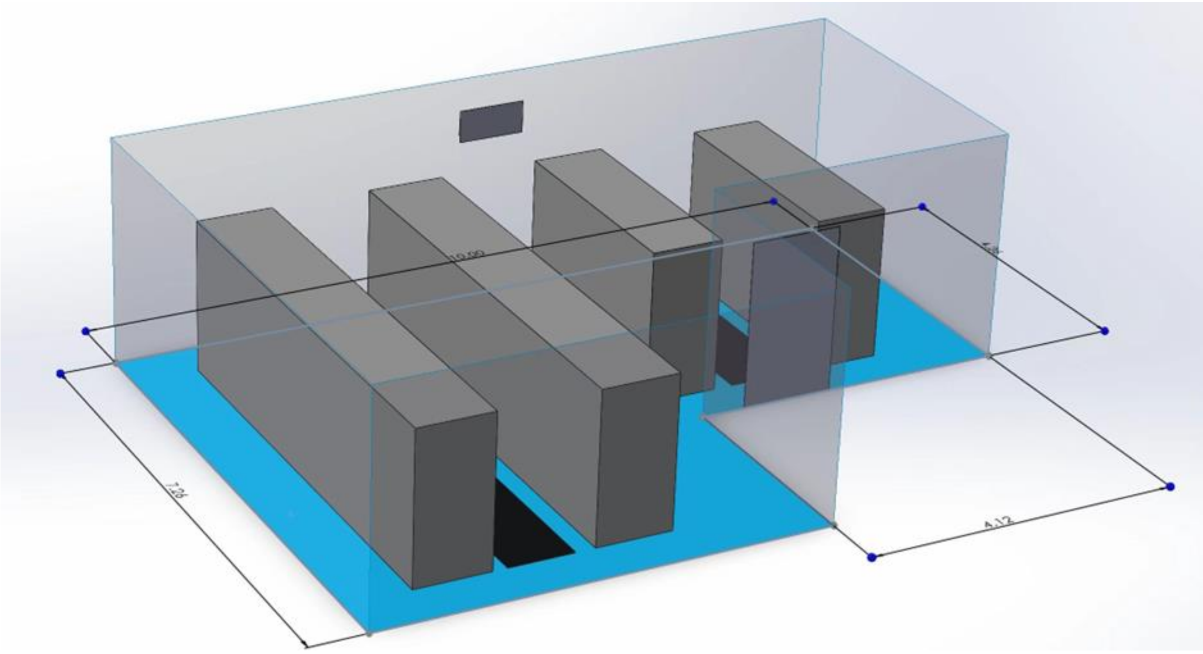


Figure 6-16 3Dview of the server room with additional servers and airway

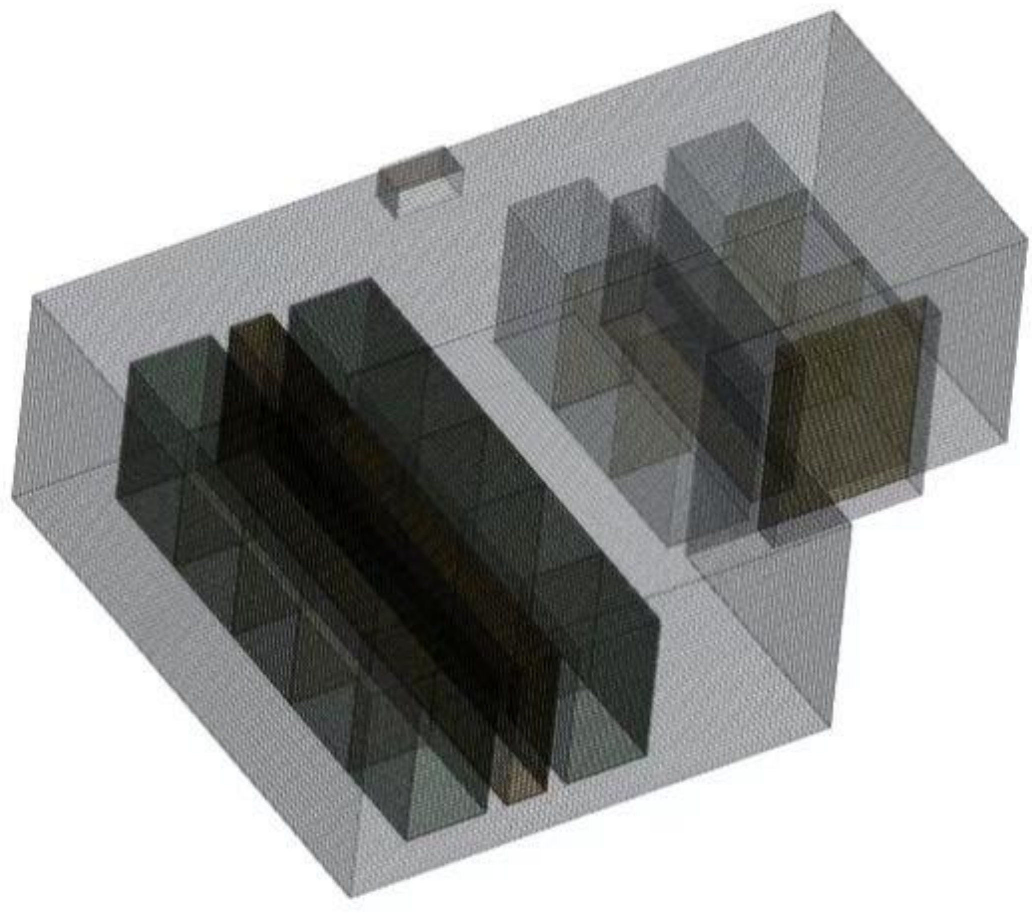


Figure 6-17 Mesh of the server room with server added.

subsequent sensitivity testing.

(2) Mesh sensitivity analysis

Table 6-3 presents the mesh quality in terms of a numerical index ranging from 0.000001 to 1, with higher values indicating better quality and lower values indicating poorer quality.

Table 6-3 Volume change statistics of the mesh

Volume change range	**Number of elements**	**Quality Index**
Volume change<1.000000-06	0	0.000%
0.000000e+00<=Volume change<1.000000-06	0	0.000%
0.000000e+00<=Volume change<1.000000-05	0	0.000%
0.000000e+00<=Volume change<1.000000-04	0	0.000%
0.000000e+00<=Volume change<1.000000-03	0	0.000%
0.000000e+00<=Volume change<1.000000-02	0	0.000%
0.000000e+00<=Volume change<1.000000-01	207	0.001%
0.000000e+00<=Volume change<1.000000-00	598928	99.965%

Additionally, the result of sensitivity analysis yielded an optimal mesh selection with a face value of 599136 and vertices of 577568, which refer to the 2D polygons that constitute the surface of the 3D object being meshed. The faces and vertices together define the geometry of the mesh. Furthermore, the mesh validity results indicated that the mesh is topologically valid, with no negative volume cells. This evaluation process ensured that the mesh was of high quality and suitable for use in the CFD simulations.

6.4.4.2 Analytical results

The current layout of the server room has been assessed to test the adaptability of the proposed control strategy in response to the modifications. The simulation temperature distribution and fluid contour have been depicted and shown in Figure 6-18, demonstrating the temperature and fluid movements when the door is appropriately closed. Due to the relatively narrow

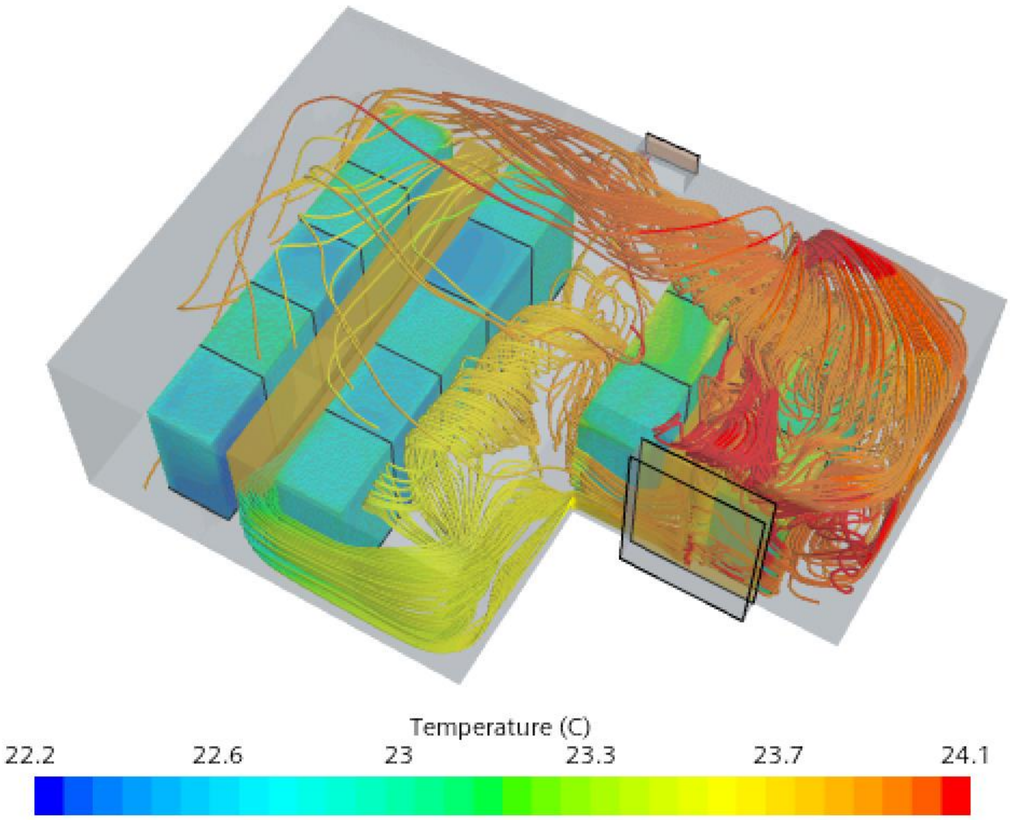


Figure 6-18 Temperature distribution and fluid contour with additional server and airway

space on the right-hand side of the server room, the temperature is comparatively higher than in other spaces in the server room.

Figure 6-19 shows the temperature changes of servers in a server room after the addition of six server racks and one cooling airway. The simulation results demonstrate that even with the increased number of servers and airflow changes, the servers' temperatures are maintained within an acceptable range during 12 busy hours of operation. The servers were operated at relatively 38% of the total load, and their temperature did not exceed the upper boundary during this period.

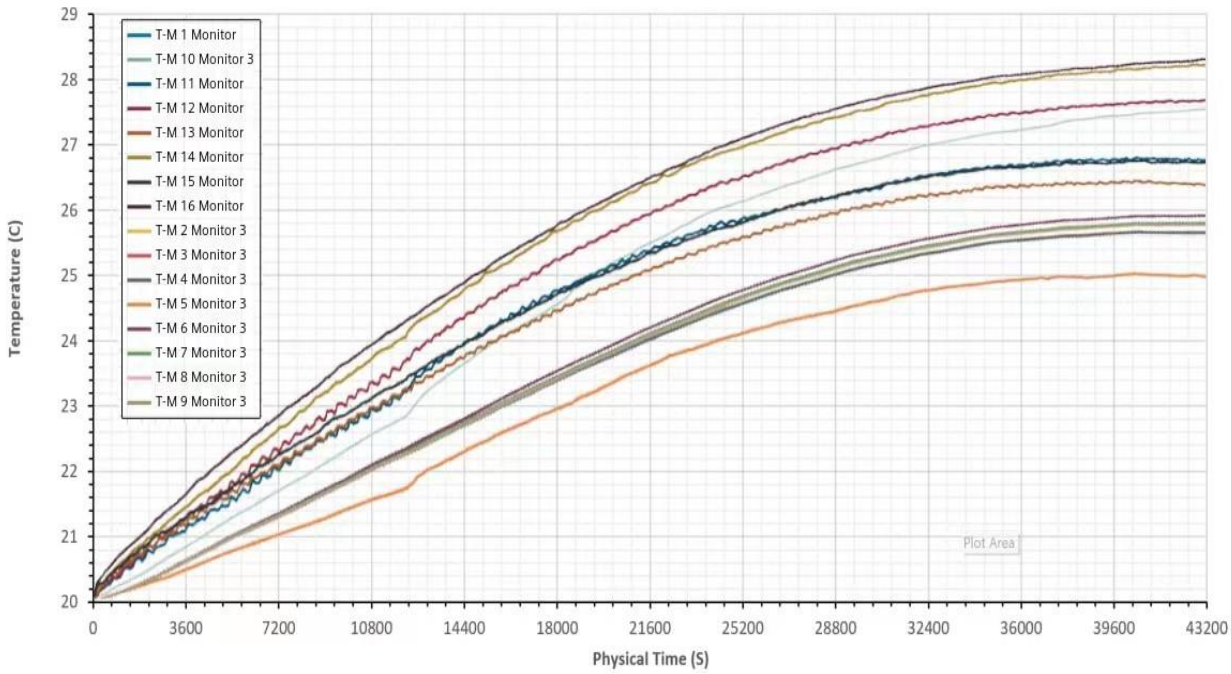


Figure 6-19 servers' temperature control after adding server racks

Figure 6-20 displays the ambient temperature of the server room after servers were added. As shown in the figure, the ambient temperature has remained below the upper boundary. By taking into account the server temperature, it can be inferred that the servers' and the room's temperatures have gradually stabilized in the last three hours.

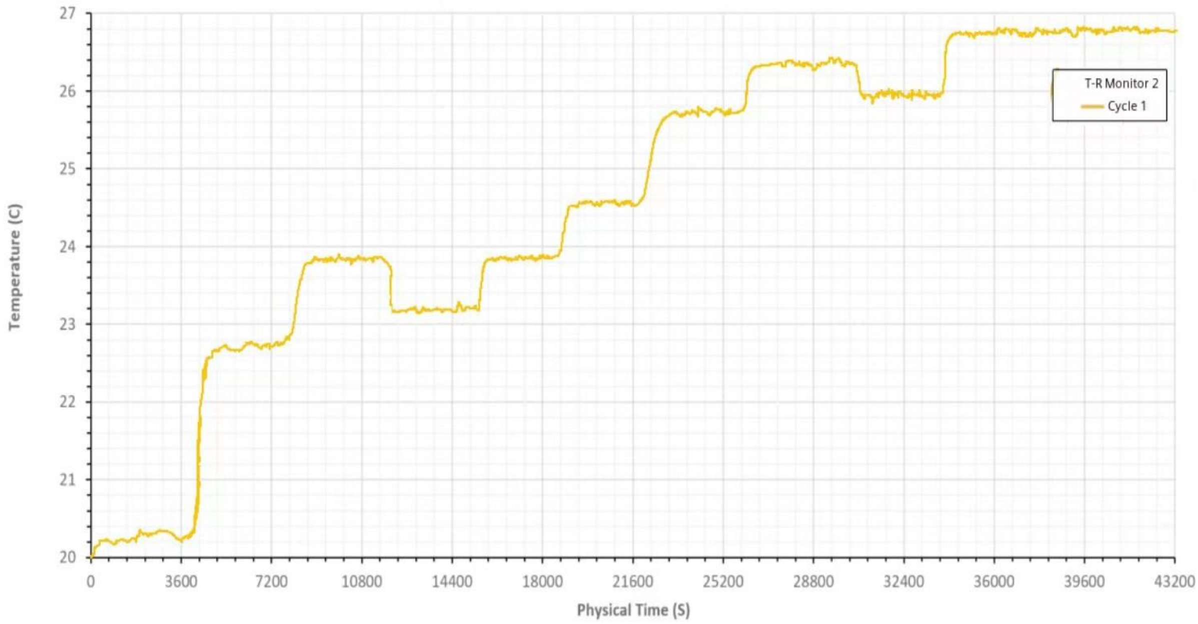


Figure 6-20 The ambient temperature after adding server racks

Figure 6-21 demonstrates the PUE movements following the adjustment of ACs temperature settings. It can be illustrated that the PUE has been significantly reduced from 2.3 to 1.6 with the temperature setpoints changes of the ACs.

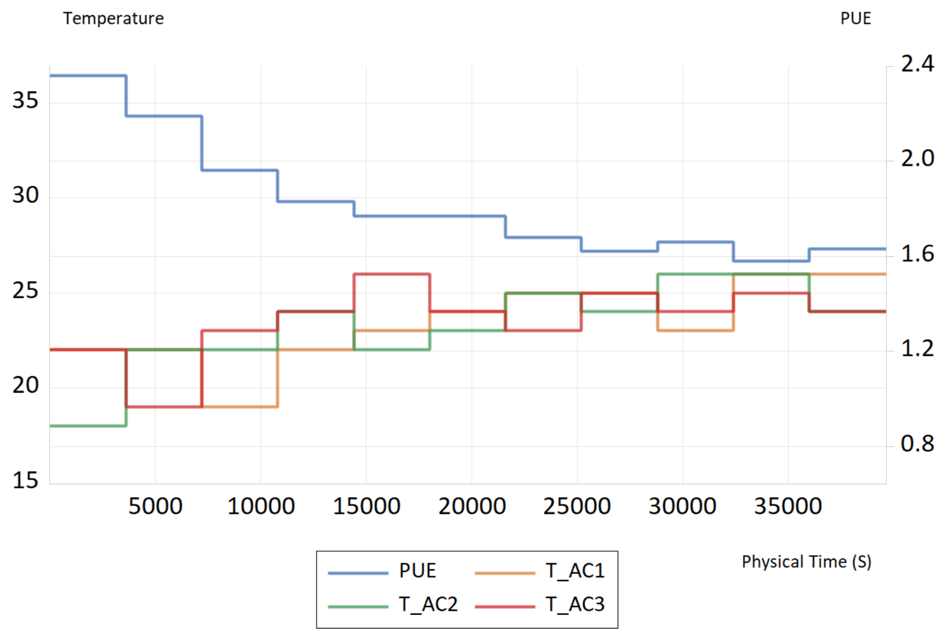


Figure 6-21 AC setpoints and PUE movements after adding server racks

6.4.5 Door-open scenario

In this section, the evaluation of the effectiveness of the control strategy in managing unforeseen occurrences in a DC is conducted through scenario testing. The focus of the testing is to investigate the impact of leaving the server room door accidentally open and its effects on the server room environment, while simultaneously assessing the response capabilities of the control strategy. The purpose of the testing is to provide a comprehensive analysis of the control strategy's capacity to maintain a stable and dependable operating environment for the servers, especially in situations where unanticipated events arise. Figure 6-22 depicts the temperature distribution and Fluid contour of the open-door scenario. The figure highlights that the fluid moves towards the door, turbulence is present, and cold air leaks from the open door, leading to energy inefficiency.

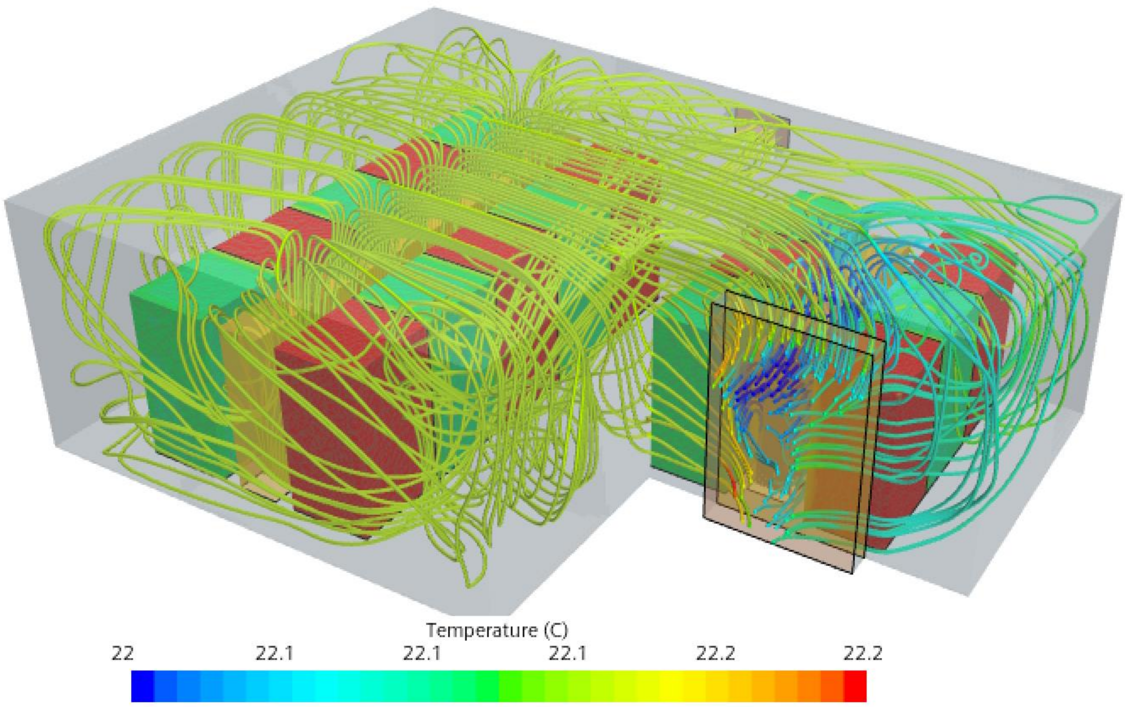


Figure 6-22 Temperature distribution and Fluid contour of the opening-door scenario

Figure 6-24 and Figure 6-23 illustrate the temperature movements of the servers and the ambient surroundings during a one-hour duration, where an accidental opening of the server room door occurred at the 1800-second mark. As evident from the figures, the server temperature experienced a sudden increase of over 0.5 ℃ within 600 seconds, while the ambient temperature underwent a significant surge of 4 ℃ within the same time frame. The implementation of the MPC control strategy facilitated the restoration of the stable temperature environment. Within 100 seconds of detecting the temperature increase, the MPC effectively curbed the upward trend, leading to a gradual decrease in server temperature in 400 seconds. The ambient temperature was also significantly reduced by 1.5 ℃ within 600 seconds, which is attributed to the effective application of the MPC control strategy.

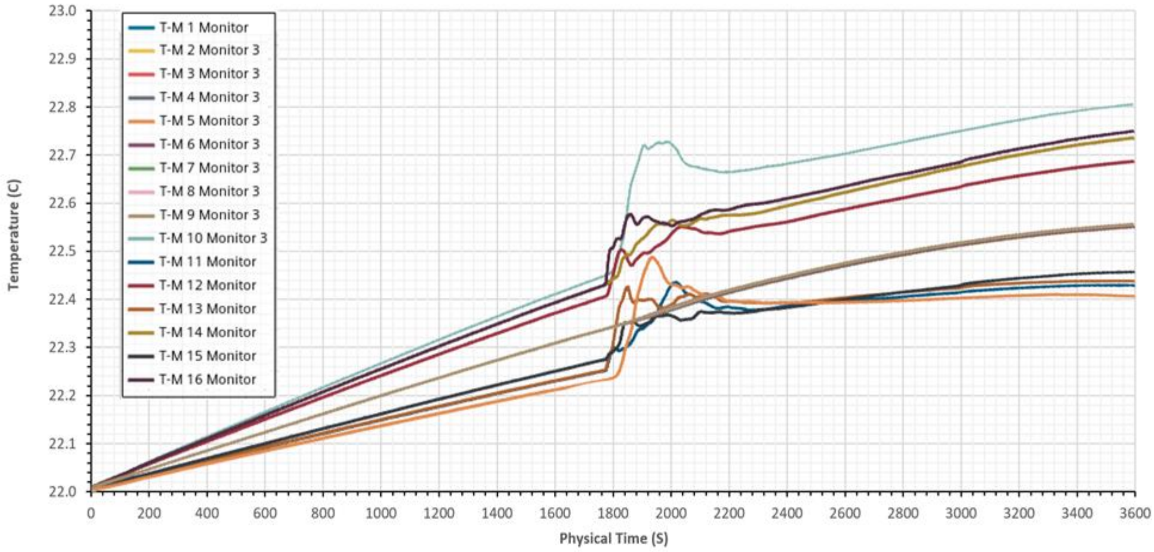


Figure 6-24 Servers temperatures under the open-door scenario

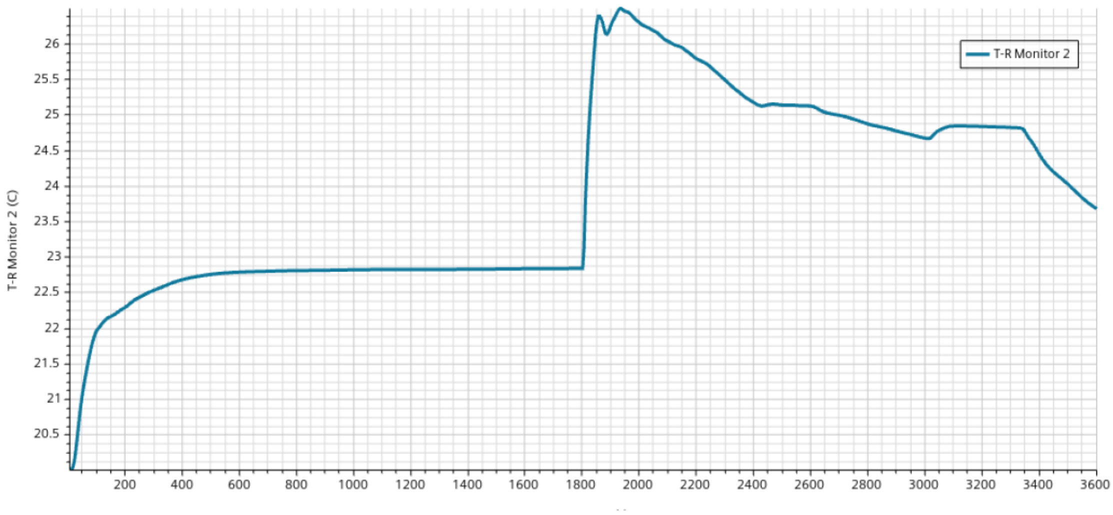


Figure 6-23 Ambient temperature under the open-door scenario

In order to enhance the control strategy, an alarming system was incorporated into the controller design. This system triggers a response when the server temperature undergoes a change of more than 0.5 ℃ within 600 seconds, or when the ambient temperature changes by more than 2 degrees without any intervention. The control system is then adjusted to impose larger penalty values on both the server and ambient temperature constraints. Additionally, the maximum velocity of the MPC control is increased to 12 degrees, which represents the largest step length, to improve the efficiency of the control.

6.4.6 Summary of Scenarios and Main Finings

In the previous sections, five scenarios have been examined under simulated environments. provides a summary of the key findings from each scenario, demonstrating the versatility of MPC control across various conditions.

Table 6-4 provides a summary of the key findings from each scenario, demonstrating the versatility of MPC control across various conditions.

Table 6-4 Summary of outcome under scenarios

Scenario Description	**Main Findings**
Scenario 1: Baseline	Stable system performance observed under normal operating conditions. Low temperature settings lead to over-cooling underscores the necessity for implementing an energy-efficient strategy.
Scenario 2: Workload Variation	The MPC control effectively regulates ambient temperature under workload variations. Despite optimisation efforts, heavily loaded servers approach upper-temperature limits when CPU usage reach 67.31%, signalling the need for additional cooling capacity.
Scenario 3: Adding Cooling Unites	Under MPC framework, adding cooling unites effectively reduced server temperatures, mitigated

	potential overheating risks, and improved PUE metrics by progressively decreasing from 2.2 to 1.7 over the course of operation.
Scenario 4: Adding Sever Racks	The addition of six server racks and one cooling airway in the server room, coupled with adjustments in AC temperature settings, maintained server temperatures within acceptable ranges during 12 busy hours of operation, stabilizing gradually, and reducing Power Usage Effectiveness (PUE) from 2.3 to 1.6, indicating a good adaption of MPC framework to the infrastructure and server room layout changes.
Scenario 5: Door-Open	The accidently opening of the server room door led to rapid increases in both server and ambient temperatures. The integration of an alarming system with heightened penalties on temperature constraints and increased maximum velocity of the MPC control significantly improved stability, swiftly mitigating temperature fluctuations within 100 seconds and gradually reducing server temperature in 400 seconds.

6.5 Adaptability evaluation of MPC control

To test the effectiveness of the MPC model in regulating the cooling system in DCs, it is crucial to evaluate its adaptive capacity in response to changes in the infrastructure, particularly those related to physical modifications of the server room. The addition of servers, for instance, can significantly alter the thermal loads and energy consumption, necessitating an adjustment in the cooling system's operation to maintain the desired temperature setpoints. The following evaluation demonstrated the performance of the proposed MPC control strategy in controlling the environment.

6.5.1 The simplicity of system model adjustment

The alterations in the number of ACs and servers, as well as modifications to the server room layout, merely lead to the inclusion of additional VAR variables. The proposed approach has the advantage of being relatively straightforward to implement. It does not require significant changes to the structure of the model, or the mathematical algorithms used to analyse the data. Instead, the new variables can be added to the existing model and the data can be re-estimated using the same techniques as before. The new data estimation and modelling cost 30 minutes in total including data collection and model adjustment. The following indices can be utilised to evaluate the VAR model's goodness of fit before and after the infrastructure changes:

- Akaike Information Criterion (AIC): The AIC is a metric used to measure the goodness of fit of a model, taking into account the number of parameters in the model. Lower values of AIC indicate a better model fit.

- Bayesian Information Criterion (BIC): Similarly, the BIC is a measure of the model's goodness of fit, however with a stronger penalty for model complexity. Lower BIC values also indicate a better model fit.

- Root Mean Squared Error (RMSE): RMSE is the square root of Mean Squared Error (MSE), providing a measure of the average deviation of the predicted values from the actual values. A lower RMSE value indicates a better model fit.

- Granger causality test: The Granger causality test is used to evaluate the causal relationships between the variables in the model. A significant Granger causality test result indicates that one variable can be used to predict another variable, which suggests a good fit for the model.

The evaluation of the model's adaptability to changes in the server room infrastructure is presented in Table 6-5. The indices used for the comparison are computed under different scenarios in terms of the infrastructure changes to assess the model's performance. The comparison of these indices provides insights into the effectiveness of the model in adapting to changes in the server room infrastructure.

Table 6-5 Goodness of fit evaluation of system model under the scenarios involving infrastructure changes

Evaluation Metric	**Baseline**	**Adding ACs**	**Adding Servers**
AIC	1246	1290	1310
BIC	1431	1450	1460
RMSE	0.043	0.038	0.039
Granger Causality	p-value=0.0327	p-value=0.0290	p-value=0.0318

The evaluation of the system model's goodness of fit under the scenarios with infrastructure changes revealed that the addition of more servers had a relatively minor impact on the model's performance. Despite a slight increase in the AIC and BIC values after the change, indicating a small reduction in model fit due to the increase in complexity, the increase of AIC and SC were not significant and can be considered acceptable. The slight decrease in RMSE after the change suggests a minor improvement in the model's predictive accuracy. Additionally, the statistically significant Granger causality test results both before and after the change indicate that the causal relationships between the variables in the model were strong in both cases. Overall, these findings suggest that adding more servers to the model was a reasonable adjustment that did not compromise the model's performance to a significant extent.

Overall, based on these results, it appears that the model was able to adapt relatively well to the changes in the server room infrastructure, with only minor changes in performance metrics observed after the addition of more servers to the model.

6.5.2 Optimisation performance evaluation

The average convergence rate and objective values during the MPC controlled period will be used to evaluate the performance of PSO optimisation under different scenarios. The comparison is shown in Table 6-6.

Table 6-6 PSO performance comparison under scenario involving infrastructure changes

Evaluation Metric	**Baseline**	**Server-load variation**	**Adding ACs**	**Adding Servers**	**Door open**

Convergence rate (number of iterations)	123	132	129	131	127
Objective function value	6.55855	6.60367	6.04120	6.03367	6.04450

The comparative results suggest that the convergence rate exhibited a slight increase following the modifications, which may indicate a longer time required for the PSO algorithm to reach an optimal solution. The server-load variation scenario resulted in a slightly higher number of iterations needed for the PSO algorithm to converge to a solution. This could indicate a slight effect of workload variation on the algorithm's performance. However, the difference in convergence rate was relatively minimal, indicating that the PSO algorithm's performance was not significantly impacted by the changes of the servers' workload.

In terms of the objective function value with the other scenarios, the results indicate a significant decrease in the objective function value under the scenarios with the additional ACs, which suggests that the PSO algorithm was more effective in minimising PUE with more sufficient cooling capacity. The objective function value also increased slightly within the acceptable range under the scenario where additional servers were added, indicating that the PSO algorithm was still able to maintain good performance even with a higher workload. Ultimately, the PSO algorithm exhibited a similar objective function value in the door open scenario as in the adding servers scenario, despite the sudden temperature increase caused by the open door. This suggests that the algorithm was able to adapt well to the changing conditions and that the MPC control was able to quickly mitigate the effect of the open door on the room temperature.

Overall, the robustness and high performance of the PSO algorithm are evident in various scenarios, such as server load variations, adding air conditioners, and adding servers. Although there are slight differences in convergence rates, the algorithm consistently minimises the PUE in all scenarios. In the door open scenario, the PSO algorithm shows excellent adaptability by promptly responding to temperature changes, and it maintains a similar objective function value to that of the adding servers scenario, demonstrating its ability to handle unexpected disturbances.

6.5.3 System running time

The results indicate that the average system running time increased from 5.21 minutes to 5.35 minutes following the modifications to the server room infrastructure. While this increase in running time may seem concerning, it is important to consider that these computations were carried out on a conventional computing device, which may not have the same computational capability as dedicated supercomputing power devices that would be used in an actual DC environment. Therefore, the current running time may be acceptable for a proof-of-concept or testing phase of the control system, as it provides an estimate of the computational resources required to run the system. Once the control system is deployed in an actual DC, more powerful computing resources could be allocated to reduce the running time. It is also worth noting that computational time is often a trade-off between accuracy and speed. In this case, a longer running time may allow for a more accurate optimisation of the AC setpoints, resulting in lower energy consumption and better control of the temperature range it is rational to expect an acceptable increase in the system's computation time as a result of the growing complexity within the system.

6.5.4 MPC control performance and stability

The performance and stability of the MPC control system in reducing PUE through temperature adjustment were evaluated under several different scenarios. Table 6-7 shows the evaluation metrics of MPC temperature control performance under different scenarios. The matrices adopted for the evaluation are as the following:

- Robustness Margin (RM): This metric is a measure of the MPC controller's ability to maintain stable control, a higer RM value indicates greater stability and robustness, meaning the controller can handle larger disturbances without deviating too far from the desired temperature setpoint.
- Standard deviation (Std.dv) under static state: This metric measures the variability of the temperature readings when the temperature is in a static state. A lower standard deviation indicates that the MPC controller can maintain more consistent and stable temperature control.
- Temperature Static Time: This metric measures the time it takes for the

temperature to reach a static state. A shorter static time indicates that the MPC controller is able to quickly respond and bring the temperature back to a stable state, indicating a faster and more effective temperature control response.

The results indicate that the MPC algorithm exhibits good control capability, as demonstrated by the consistent RM values across all scenarios, The addition of ACs has the greatest positive impact on the controller's ability to maintain stable control. However, the stability margin is slightly reduced when additional servers are added, likely due to the increased complexity of the system. In addition, the scenario with server workload variation has presented increased difficulties for the controllers to maintain the desired temperatures However, the Robustness Margin (RM) values for this scenario still fall within an acceptable range, indicating that the controllers have a good level of adaptability to handle varying workload conditions. Moreover, it can be observed that the standard deviation (Std.dv) values under static state are relatively low, suggesting that the temperature control is stable across all scenarios, with the exception of a slight variation observed in the server-load variation scenario. Furthermore, the table also presents the temperature static time (S) for each scenario, which provide an indication of the time required for the temperature to reach a steady state. It can be observed that the time consuming are relatively short for all the scenarios, indicating that the MPC control is able to quickly adjust to changes in the system and maintain stable temperature control.

Table 6-7 Performance evaluation of MPC temperature control

Metrics	Baseline	Server-load variation	Adding ACs	Adding servers	Open door
Robustness Margin (RM)	+/-0.88	+/-0.79	+/-0.92	+/-0.84	+/-0.76
Std.dv under static state	0.433	0.528	0.397	0.416	0.407
Temperature Static Time (S)	2538	2670	2321	2572	2651

CHAPTER 7

RESEARCH FINDINGS AND DISCUSSION

In this final chapter, the emphasis is on evaluating the performance of the proposed framework for optimal cooling system control in data centres (DCs), as well as summarizing the methodological advantages of the approach based on empirical evidence. The chapter provides a comprehensive overview of the research findings, which include results from both field experiments and simulations. Furthermore, the chapter examines the potential implications of the proposed framework for the DCs industry and discusses possible future directions for further research.

7.1Problems restatement and contributions of the study

Data centres (DCs) require a significant amount of energy to maintain the optimal environmental conditions for their servers, resulting in a considerable impact on operational costs and the environment. The Power Usage Effectiveness (PUE) metric has been established to evaluate DCs efficiency, and a lower PUE signifies a more efficient operation, leading to cost savings and environmental benefits. As such, reducing the PUE has emerged as a crucial challenge for DCs operators. While several techniques, such as cooling optimisation, air management, and equipment upgrades, have been proposed to reduce PUE, their implementation is often static and does not consider the dynamic relationships among variables in the server room. Moreover, these techniques may require significant investments in new equipment or infrastructure, thereby limiting their accessibility to the DCs, especially small and medium-sized ones.

One of the main challenges in managing DCs is to maintain a low PUE value while ensuring that the servers are running efficiently and reliably. This requires a careful balance of cooling and power supply systems to match the IT load and minimise energy waste. To address the problem of energy waste in DCs, a managerial solution is needed that optimises energy consumption while maintaining the required level of service quality.

The proposed VAR-PSO-MPC framework presents a novel solution to the challenge of achieving optimal cooling system control in DCs without requiring significant investments in new equipment or infrastructure. The integration of VAR modelling, PSO optimisation, and MPC control in the framework was found to be crucial for achieving optimal cooling system control in DCs. The VAR model was able to capture the dynamic relationships among the variables in the server room, while the PSO algorithm was able to optimise the setpoints for the air conditioners in real-time. The MPC control framework was able to use the optimised setpoints to achieve the desired cooling system performance while maintaining the server inlet temperature within the desired range. Through the integration of the three techniques, the framework can address the dynamic relationships among the variables in the server room and achieve optimal control of the cooling system.

Moreover, the flexibility and adaptability of the proposed control framework enable it to be customised to different DC environments, without adding additional costs for renewing the facilities. This makes it a cost-effective solution for improving energy efficiency and reducing operating costs in existing DCs. By providing DC operators with an accessible and practical framework for optimizing their cooling systems, this approach has the potential to significantly reduce energy consumption and environmental impact in the DCs industry.

7.2 Summary of research findings

The proposed VAR-PSO-MPC framework was tested through a combination of field experiments and simulation studies. The field experiment was conducted in a small-sized DC over 4 weeks, and the simulation studies were conducted based on an environment built by Computational Fluid Dynamic (CFD) simulation tool. The results of the study are summarised as follows:

(1) The VAR-PSO-MPC framework was proven effective in decreasing PUE in both field experiments and simulation studies. The framework resulted in an average reduction of PUE of 15.57% (from 2.20 to 1.86) compared to the baseline scenario in the field experiment, while in simulation studies, it achieved an average reduction of PUE of 17.39% (from 2.30 to 1.90) over 3 hours with constant server workload and a reduction of 25% (from 1.90 to 1.42) over 8 hours with varying server workload.

(2) The VAR-PSO-MPC framework is highly effective in maintaining optimal server performance and preventing hardware failures by regulating the server inlet temperature within the desired range. In the field experiment, the framework maintained the server inlet temperature within the desired range at 100% of the time, compared to 93.7% for the baseline scenario. Additionally, during simulation studies, the framework was able to keep the server inlet temperature below the upper boundary of 32°C, provided that the cooling capacity was sufficient.

(3) The experiment is also successful in predicting and avoiding over-heated. The result suggested that the baseline capacity of cooling is insufficient when the workload exceeds 67%, additional cooling capacity would be required. Illustrated the crucial role of the control framework in maintaining optimal server performance and preventing hardware failures.

(4) The field experiment provided evidence that the implementation of the suggested control framework has resulted in a significant reduction in

power consumption and greenhouse gas emissions, demonstrating a remarkable positive environmental impact. When comparing the before and after scenarios, a significant reduction of 20.65 kWh, equivalent to a 32.5% decrease in energy consumption, was observed. This resulted in a potential annual energy saving of 307,505 kW and a cost reduction of 1,537,524 THB (equivalent to 35,982 GBP). In addition, the decrease in power consumption contributed to a reduction of 405,906 lbs of greenhouse gas emissions, indicating a positive impact on the environment.

(5) The results of the study also suggested that increasing the frequency of control would result in an even better performance of the control framework. The study showed that the VAR-PSO-MPC framework was effective in regulating the server inlet temperature and reducing PUE, but a higher frequency of control could further optimise the performance of the system. By increasing the frequency of control, the framework would be able to respond to changes more quickly in server workload and environmental conditions, as well as maintain optimal performance with greater accuracy. Therefore, implementing a more frequent control schedule would be beneficial in maximising the efficiency and effectiveness of the VAR-PSO-MPC framework.

(6) The VAR-PSO-MPC control framework demonstrated its ability to adapt to changes in the infrastructure, including modifications to the cooling system or server configuration. For instance, the framework automatically adjusted the cooling system to maintain optimal performance when a new server was added to the system. In addition,

the framework responded swiftly to unexpected conditions, such as an open door or a cooling system malfunction. The control framework detected temperature changes and compensated for heat gain by increasing cooling capacity in the event of an open door. Furthermore, in case of a cooling system malfunction, the framework promptly notified the maintenance team to prevent a significant impact on server performance.

In conclusion, the VAR-PSO-MPC control framework can bring significant benefits in terms of energy efficiency, cost reduction, and environmental impact, making it a promising solution for DCs operators looking to improve their operations.

7.3 Discussion based on empirical studies

This study proposes operational strategies for reducing energy consumption in DCs, focusing on managerial strategies rather than hardware innovations. Table 7-1 presents empirical studies grouped into three main categories: cooling configuration design, IT design, and thermal management, with this study introducing a new category focusing on operational management presented in the end as the fourth category.

Table 7-1 Existing approaches to energy efficient data centres

Area	**Exemplary Approaches**	**Reference**
Cooling configuration design	A free cooling technology	Daraghmeh and Wang (2017)
	Liquid cooling technology	Carbó et al. (2016b)
	The two-phase flow technology	Riofrío et al. (2016)
	The building envelope technology	Akeiber et al. (2016)
IT design	Server consolidation	Verma et al. (2009)
	Server virtualization	Schulz (2011)
	Storage consolidation	Zhang et al. (2018)
	Decommissioning the idle servers	Pöyhönen et al. (2021)
Thermal management	Adapting the server racks to the hot /cold aisle layout as well as the containment and enclosure methods	(Niemann et al. 2013)
	The optimal layouts of the cooling devices according to the thermal dynamics	(Stahl & Sullivan 2001),(Patel et al. 2002)

Operational Management	Implemented managerial strategies to decrease energy consumption in the DCs by actively adjusting ACs' temperature settings, without necessitating any infrastructure alterations.	This Study

Although there are remarkable breakthroughs in the field of the DCs' hardware innovations, one of the common disadvantages of the existing approaches is the complexity and high cost of upgrading and replacing the existing infrastructure. Existing new technologies require additional time to verify, improve and mature before applying to the practical DCs. The high cost of replacing the existing facilities is also crucial for realising the energy-saving goal of the DCs through hardware renovations and replacement. The report from Forbes Technology Council (R. Danilak 2020) evaluated that would be hardly seen some essential improvements in the above areas in the next five years since a benefit from the adoption of new energy-efficient techniques for DCs would be easily compromised by the cost of renewing the DC infrastructures, this essentially became one of the barriers to realising the facilities renovations. Therefore, this circumstance sparks our motivation to investigate managerial strategies instead of changing the existing facilities in DC to reduce energy cost.

By adopting the proposed MPC framework, the DCs can optimise their operational efficiency without the need for hardware modifications. Although the general advantages of MPC are significant, we have taken the additional step of addressing some of its limitations, thereby filling a crucial gap in the technology's capabilities. Some academics argued that the additional computing load brought by the design cannot be overlooked even though the system performance is rigorously guaranteed by theory (XI et al. 2013). This is also the main reason that the application field questioned the feasibility of MPC theories. In response to this issue, we adopted the technique of "offline design, online synthesis" in the qualitative synthesis of predictive control. By converting part of the online calculation of the integrated control law into the offline calculation, the purpose of reducing the amount of online calculation is achieved. The joint implementation of offline and online largely saved computational time and effort, allowing the proposed MPC framework to be applied in a real-time industrial environment.

Furthermore, the general MPC requires plant models, to deeply understand and model the nature of the environment. General plant models that are based on the first principles (physics-based models) can be laborious and computationally expensive due to the complexity of the environment (Behrooz et al., 2018). There is a wide range of literature that provides additional insight into the development of first principles models. Take an example of the work conducted by Rehrl and Horn (2011) on HAVC temperature control systems, several separated plant models have been derived from the thermal dynamic principles. To simulate the plant dynamics, physics-based models for the valve gear, hydraulics, cooling coil, and temperature sensor have been developed. Although the results show a good performance index with the adoption of MPC controller and the feedback linearisation, the substantial computational cost and the effort required to fully comprehend the characteristics of the system devices cannot be overlooked. In comparison, data-driven models have higher precision than physics-based models (Perera, Pfeiffer and Skeie, 2014). In this study, the effort of studying the plant has been avoided and the adoption of the data-driven model VAR ensures the nature of the environment can be captured in real-time. Due to the benefits of VAR, we successfully handled the multiple control variables and the large size of the parameter, the disadvantage of large computational load in general MPC industrial examples has been largely reduced in our study.

Moreover, the approach proposed in this study not only offers an energy-efficient solution and addresses the limitations of conventional MPC, but also holds significant implications for the field of optimisation. As the magnitude of data increases and hardware capabilities improve, optimisation issues tend to be tied to more and more scenarios, and the dimensions of the algorithms grow proportionately. One way to solve such high-dimensional data is to impose certain structural constraints on the problem from the perspective of parameter estimation, which is often non-convex. Correspondingly, the objective functions of such optimisation problems are also non-convex. Non-convex objective functions and constraints can model the problems more accurately, however tackling such problems might be difficult. Solving the objective function by traditional mathematical programming approach is mostly NP-hard, and addressing approximation solutions is also likewise NP-hard (Jain & Kar 2017). Academic research demonstrates that solving optimisation problems in high dimensions (large candidate solutions) can be challenging because it is time-consuming and computationally expensive during calculation by mathematical evolution algorithms (Tomassetti & Cagnina 2013) and nonlinear algorithms (Du & Chen 2000). R.Urbanucci (2018) categorised the optimisation solutions for energy polygeneration systems into three types, including optimum synthesis, design, and operations, in which MILP (Mixed Integer Linear Programming) is the approach for operations with the main advantage in finding the global optimal solution to the problem and easy to be solved by many commercial solvers. However, in the complex industrial practice, MILP also has limitations. D.Steen (2015) derived a thermal storage model for energy-saving by MILP and confirmed that because of the existence of the endogenous problem in the thermal environment, the temperature storage cannot be tracked by MILP for the whole timeline horizon but only a single time step. To address these issues, our proposed method integrates endogenous variables and exogenous variables by the VAR model. In a real-time industrial circumstance, the data-driven model is more flexible to adapt to the changing environment compared to general mathematical models. Additionally, the MPC control strategy has broken down complex high-dimensional optimisation problems into linear problems with a shorter time horizon. Because of the finite control horizon and a fixed moving window, linear control problems under MPC control are easy to be transferred to quadratic programming problems, which is more computationally efficient(Rao et al. 2001).

The following context summarises the key advantages over other similar studies in optimising cooling systems in DCs. This section summarises these advantages and compares them to other existing studies.

(1) Real-Time Adaptability

Based on the comparison to similar studies in the literature, the proposed VAR-PSO-MPC approach offers a significant advantage over traditional control methods and other MPC-based approaches in terms of real-time adaptability. Specifically, the use of a VAR model and PSO optimisation enables the system to quickly adapt to changes in server workload and environmental conditions, leading to improved system performance and energy efficiency. In comparison, traditional control methods such as On/Off control and PID control rely on predetermined setpoints and fail to adapt to changing conditions, resulting in unnecessary energy consumption and decreased system performance. Additionally, similar studies that rely on linear models and predetermined setpoints, such as the approach by Liu et al. (2016), are also limited in their real-time adaptability when compared to the proposed VAR-PSO-MPC approach.

(2) Data-driven characteristic

Another advantage of the proposed approach is its data-driven characteristic. The VAR model used in the approach is based on machine learning techniques and uses historical data to predict future system behaviour. This data-driven approach allows for more accurate predictions and more effective control inputs, leading to improved energy efficiency and system performance. In contrast, traditional control methods such as On/Off control and PID control rely on pre-determined setpoints and do not consider the dynamic relationship between the cooling system and server performance. Zhao et al. (2020) also employed machine learning techniques to optimise cooling systems in DCs. However, their method utilizes a combination of a neural network (NN) and genetic algorithm (GA), which differs from the VAR-PSO-MPC approach proposed in this study. In comparison, the proposed approach demonstrates a more data-driven strategy as it uses historical data to forecast future system behaviour through the VAR model. This enhances the accuracy of predictions and control inputs, thus leading to improved energy efficiency and system performance. In contrast, the neural network employed by Zhao et al. (2020). is only trained on present data, which may limit its ability to forecast future behaviour effectively. Moreover, the GA algorithm applied in Zhao et al. (2020)'s. study may not be as proficient in finding optimal control inputs as the PSO optimisation used in the proposed approach.

(3) Nonlinear Problem Handling

A comparable study has been conducted by Li et al. (2020) who proposed an optimal control approach for dynamic cooling management in DCs that utilizes MPC control. However, their approach does not explicitly consider the non-linear relationship between the cooling system and server performance. In comparison, the proposed VAR-PSO-MPC approach integrates a VAR model-based approach capable of handling non-linear problems. The VAR model allows for the possibility of lagged variables influencing the current state of the system, capturing dynamic relationships between variables, including any non-linear relationships that may exist. Therefore, the proposed approach accounts for the complex interactions between the cooling system and server performance, resulting in improved energy efficiency and system performance.

Moreover, It is worth noting that traditional control methods, such as PID control or On/Off control, rely on linear models and may not be effective in handling non-linear relationships between variables. This highlights the superiority of the proposed VAR-PSO-MPC approach, which incorporates a VAR model-based approach capable of handling non-linear problems by capturing dynamic relationships between variables, including any non-linear relationships that may exist. Consequently, the proposed approach can provide more accurate predictions and control inputs, leading to improved energy efficiency and system performance in complex DC cooling systems.

(4) The use of PSO optimisation

The use of PSO optimisation in the proposed VAR-PSO-MPC approach is a notable advantage over other comparable studies. This optimisation technique allows for the efficient identification of optimal control inputs, resulting in improved energy efficiency and system performance. The effectiveness of PSO optimisation can also be attributed to the incorporation of the VAR model, which enables accurate predictions of the system behaviour. By utilising the VAR model in the proposed VAR-PSO-MPC approach, PSO optimisation is more effective in identifying optimal control inputs, leading to improved energy efficiency and system performance. In contrast, many existing studies rely on traditional optimisation techniques such as linear programming, which can be time-consuming and less effective in complex DC cooling systems. One such comparable study proposed by Heidarzadeh et al. (2018) also utilises MPC control in DC cooling systems. However, this study employs a modified genetic algorithm (GA) to optimise control inputs, which may not be as effective as

PSO optimisation. Genetic algorithms may struggle to find optimal control inputs in complex DC cooling systems, leading to reduced energy efficiency and system performance. In contrast, PSO optimisation is based on swarm intelligence, allowing multiple particles to simultaneously search the solution space for faster and more accurate identification of optimal control inputs.

(5) Greenhouse Gas Emissions Reduction

The proposed VAR-PSO-MPC approach has shown significant reductions in greenhouse gas emissions, which is another advantage over other similar studies. The approach achieves these reductions by maintaining optimal server performance and preventing hardware failures. This capability leads to improved energy efficiency and reduced greenhouse gas emissions in DC cooling systems. A comparable study to our proposed VAR-PSO-MPC approach is conducted by Ma et al. (2018). However, while this study also employs MPC control in DC cooling systems, it does not explicitly focus on addressing the environmental impact of DC operations. The proposed VAR-PSO-MPC approach considers the dynamic relationship between the cooling system and server performance, leading to improved energy efficiency and reduced greenhouse gas emissions, while Ma (2018) focused solely on temperature control in DCs without considering the impact on server performance, potentially resulting in suboptimal energy efficiency and higher greenhouse gas emissions.

(6) Comprehensive Approach

The proposed VAR-PSO-MPC approach is a comprehensive approach to handling complex DC cooling systems. The joint use of VAR, PSO optimisation, and MPC control allows for the advantages of each method to be leveraged, providing a more comprehensive approach to handling non-linear complex problems. In contrast, many existing studies have relied on a single control method, such as MPC or On/Off control, leading to suboptimal energy efficiency and system performance. A comparable study by Mohan et al. (2016) also utilises MPC control in DC cooling systems, but it only focuses on optimising the cooling system's performance and does not consider the overall system's energy efficiency. In contrast, the study by Mohan et al. (2016) solely relies on a thermal model and a deterministic algorithm that utilises linear algebra to iteratively enhance a single solution, without incorporating any dynamic system model or advanced optimisation techniques to further enhance the control's performance. This approach may be less efficient compared to the proposed VAR-PSO-MPC approach, which utilises a combination of advanced techniques to handle non-linear complex problems in DC cooling systems. Hence, the proposed VAR-PSO-MPC approach offers a more comprehensive and effective solution for handling complex DC cooling systems when compared to existing studies that rely on a single control method.

7.4 Managerial implications

Motivated by the discussions surrounding the critical energy consumption in DCs and recognising the existing gaps in current energy-efficient solutions, our research is dedicated to advancing the discourse on energy efficient DCs across three pivotal dimensions: theoretically, methodologically, and practically.

In a theoretical context, our focus lies in developing a managerial strategy to minimise energy consumption within DCs, particularly targeting the predominant energy usage by cooling devices. This involves an active intervention in the ACs temperature setpoints, presenting a cost-effective alternative to extensive hardware innovations.

Methodologically, our approach offers a flexible, real-time adaptable, and computationally inexpensive solution, contrasting with existing methods that require a deep understanding of system physics, extensive design efforts, and high computational costs. By introducing this methodology, we have paved the way for an efficient industry management solution that can be implemented with greater ease and at reduced costs compared to traditional approaches. Most importantly, its adaptability, scalability and flexibility ensure that a broader adoption of the proposed methodology to other complex industrial practices.

In practical terms, the proposed framework has achieved a reduction in energy costs to a desired level and ensured the safe operation of devices within the DC environment. Moreover, the potentially saved budget has been rescheduled to be used to innovate the hardware of the DC. By applying managerial strategies, the DC can reduce the significant cost of cooling energy consumption. This opens up the possibility of upgrading the hardware and implementing an advanced control system, resulting in more efficient operations and management. Furthermore, the positive result in energy saving of the DC encourages the relevant personnel to pursue green operations. Due to its tropical location and year-round low-temperature control, the data centre uses a lot of electrical power in cooling. Our proposed experiment shows considerable energy-saving potential in the DC, this encourages the DCs to pursue energy-saving goals by controlling the AC's temperature setpoints rather than changing the hardware of the DC.

Notably, the statistics arouse the environmental awareness of global DC operators. The Proposed experiment shows the potential of global green gas reduction of 405,906 lbs, which is a substantial amount that addresses the importance of environmental awareness and sustainable green operations in the DC industries.

7.5 Limitations and future direction

This section aims to discuss the limitations of the present study and highlight potential avenues for future research.

(1) There is a possibility of reducing the implementation time by simplifying the algorithm, slowing down the weight update rules, adjusting the parameter weights, reducing the size of input data, and other means. However, further experimentation is required to validate these approaches.

(2) Further research is necessary to explore the applicability of the proposed methodology to various types of DCs and cooling systems, such as water cooling, liquid immersion cooling, chilled water systems and other economisers. The generalisability of the approach should be thoroughly investigated to provide a comprehensive understanding of

its potential benefits and limitations in diverse contexts.

(3) In the field experiment, the installation of temperature sensors on every server is not feasible due to the limited number of sensors available. Thus, it is highly recommended that DCs closely monitor the inlet temperature of all servers to improve accuracy control and prevent server overheating. To further enhance the accuracy and precision of the control system, it is advisable to incorporate more advanced sensor technology and data analytics.

(4) In the simulation environment, the feasibility of the MPC control was tested by manually adjusting the server workload in the CFD simulation environment. While it demonstrated good performance in maintaining temperature and PUE, it is anticipated that applying an AI-controlled workload variation to the simulation environment would further test the response speed and performance of the control framework.

(5) To enhance the effectiveness of the control framework and lower computational costs, future research directions will be directed towards accuracy control and exploring the use of servers' inlet temperature alarm mechanics. Achieving more precise control over the cooling system while reducing computational costs would require the development of advanced control algorithms that can process large amounts of data in real-time.

(6) As a potential avenue for future research, it is suggested that the proposed MPC control framework be expanded to incorporate more

variables while maintaining computational efficiency and model fitness, in response to the increased usage of advanced sensor technology and additional sensor networks. Additionally, it is imperative to investigate the feasibility of implementing a closed-loop MPC control system that incorporates direct feedback from sensors to further enhance system control and optimisation. An improvement in the efficiency, accuracy, and robustness of the control framework is highly anticipated in the future.

(7) Future studies can investigate the impact of workload allocation on the performance of the MPC control framework. This can include exploring the optimal allocation of workloads to servers based on their capacity, utilisation, and energy consumption. Additionally, research can be conducted on optimising the fan speed of the servers to further improve cooling efficiency and reduce energy consumption.

(8) To achieve more precise density control, the MPC control system should be further enhanced to operate as an AI control system instead of relying solely on manual adjustments. Moreover, the development of better computational capabilities would be crucial in achieving more effective control. Further testing and validation of the proposed methodology in industrial settings are also necessary to assess its effectiveness and potential for improvement.

REFERENCES

Afshinmanesh, F., Marandi, A., & Rahimi-Kian, A. (2005). A novel binary particle swarm optimisation method using artificial immune system. *EUROCON 2005 - The International Conference on Computer as a Tool*, **I**(February), 217–220.

Agrawal, M. (2020). What is Model Predictive Control (MPC)? *Moor's Lobby Podcast*.

Akeiber, H., Nejat, P., Majid, M. Z. A., … Zaki, S. A. (2016). A review on phase change material (PCM) for sustainable passive cooling in building envelopes. *Renewable and Sustainable Energy Reviews*, **60**, 1470–1497.

Ang, K. H., Chong, G., & Li, Y. (2005). PID control system analysis, design, and technology. *IEEE Transactions on Control Systems Technology*, **13**(4), 559–576.

Athavale, J., Yoda, M., & Joshi, Y. (2019). Comparison of data driven modeling approaches for temperature prediction in data centres. *International Journal of Heat and Mass Transfer*, **135**, 1039–1052.

Biegler, Lorenz, et al. *Large-Scale Optimization with Applications*. 1997.

B . Lee and L . Markus. (2021). Review Reviewed Work Foundations of Optimal Control. *Journal of the Royal Statistical Society*, **132**(1), 3–4.

Bastkowski, S., Moulton, V., Spillner, A., & Wu, T. (2016). The minimum evolution problem is hard: a link between tree inference and graph clustering problems. *Bioinformatics*, **32**(4), 518–522.

Beni, G., & Wang, J. (1993). Swarm Intelligence in Cellular Robotic Systems. *Robots and Biological Systems: Towards a New Bionics?*, (2), 703–712.

Boyd, S. (2004). *Convex Optimisation*.

Calude, C. S., Dinneen, M. J., Păun, G., Pérez Jiménez, M. J., & Rozenberg, G. (2005). *Lecture Notes in Computer Science: Preface*. *Lecture Notes in Computer Science*, Vol. 3699.

Carbó, A., Oró, E., Salom, J., Canuto, M., MacÍas, M., & Guitart, J. (2016a). Experimental and numerical analysis for potential heat reuse in liquid cooled data centres. *Energy Conversion and Management*, **112**, 135–145.

Carbó, A., Oró, E., Salom, J., Canuto, M., MacÍas, M., & Guitart, J. (2016b). Experimental and numerical analysis for potential heat reuse in liquid cooled data centres. *Energy Conversion and Management*, **112**, 135–145.

Chemical Engineering Department King Saud University. (2002). (1) (PDF) Process Control in the Chemical Industries MODEL PREDICTIVE CONTROL | Youssef Nouinou - Academia.edu. Retrieved May 11, 2021, from https://www.academia.edu/30145893/Process_Control_in_the_Chemical_Industries_MODEL_PREDICTIVE_CONTROL

Cheung, H., Wang, S., Zhuang, C., & Gu, J. (2018). A simplified power consumption model of information technology (IT) equipment in data centres for energy system real-time dynamic simulation. *Applied Energy*, **222**(March), 329–342.

Daraghmeh, H. M., & Wang, C. (2017). A review of current status of free cooling in datacenters. *Applied Thermal Engineering*, **114**, 1224–1239.

Du, X., & Chen, W. (2000). Towards a better understanding of modeling feasibility robustness in engineering design. *Journal of Mechanical Design, Transactions of the ASME*, **122**(4), 385–394.

Dufour, P., Michaud, D. J., Touré, Y., & Dhurjati, P. S. (2004). A partial differential equation model predictive control strategy: Application to autoclave composite processing. *Computers and Chemical Engineering*, **28**(4), 545–556.

Eason, E. G., Carver, N. S., Kelty-Stephen, D. G., & Fausto-Sterling, A. (2020). Using Vector Autoregression Modeling to Reveal Bidirectional Relationships in Gender/Sex-Related Interactions in Mother–Infant Dyads. *Frontiers in Psychology*, **11**(August), 1–19.

Emerson Network Power Energy Logic 2.0 New Strategies for Cutting Data centre. (2012). *Emerson Network Power*, p. 8.

Feng, Y., Wang, G. G., Deb, S., Lu, M., & Zhao, X. J. (2017). Solving 0–1 knapsack problem by a novel binary monarch butterfly optimisation. *Neural Computing and Applications*, **28**(7), 1619–1634.

Ferreira, A. M., & Pernici, B. (2016). *Managing the complex data centre environment: an Integrated Energy-aware Framework*. *Computing*, Vol. 98, Springer Vienna. doi:10.1007/s00607-014-0405-x

Friedman, J., Hastie, T., & Tibshirani, R. (2000, April 1). Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, Institute of Mathematical Statistics, pp. 337–407.

Gros, S., Zanon, M., Quirynen, R., Bemporad, A., & Diehl, M. (2020). From linear to nonlinear MPC: bridging the gap via the real-time iteration. *International Journal of Control*, **93**(1), 62–80.

Hevner, Alan R., and Samir. Chatterjee. *Design Research in Information Systems : Theory and Practice*. Springer, 2010, p. 320.

Ham, S., Park, J., & Jeong, J. (2015). Optimum supply air temperature ranges of various air-side economizers in a modular data centre. *Applied Thermal Engineering*, **77**, 163–179.

Hassan Khooban, M., Vafamand, N., Niknam, T., Dragicevic, T., & Blaabjerg, F. (2016). Model-predictive control based on Takagi-Sugeno fuzzy model for electrical vehicles delayed model. doi:10.1049/iet-epa.2016.0508

Hillerström, G., & Walgama, K. (1996). Repetitive Control Theory and Applications - A Survey. *IFAC Proceedings Volumes*, **29**(1), 1446–1451.

Holiš, R., & Bobál, V. (2015). Model predictive control of time-delay systems with measurable disturbance compensation. *Proceedings of the 2015 20th International Conference on Process Control, PC 2015*, **2015-July**, 209–214.

Houwing, M., Negenborn, R. R., & de Schutter, B. (2008). Economic advantages of applying model predictive control to distributed energy resources: The case of micro-CHP systems. *2008 Mediterranean Conference* *on Control and Automation - Conference Proceedings, MED'08*, **73**, 1550–1555.

Hu, J., & Ding, B. (2018). An efficient offline implementation for output feedback min-max MPC. doi:10.1002/rnc.4401

Hu, X., Eberhart, R. C., & Shi, Y. (2003). Swarm intelligence for permutation optimisation: A case study of n-queens problem. *2003 IEEE Swarm Intelligence Symposium, SIS 2003 - Proceedings*, (February 2015), 243–246.

Heidarzadeh, M., Alavi, S. E., & Gharehpetian, G. B. (2018). Optimal Control of Data Center Cooling Systems using MPC and Modified GA. Journal of Thermal Science and Engineering Applications, 10(6), 061013.

Izawa, A., & Fripp, M. (2018). Multi-objective control of air conditioning improves cost, comfort and system energy balance. *Energies*, **11**(9), 1–18.

Jain, P., & Kar, P. (2017). Non-convex optimisation for machine learning. *Foundations and Trends in Machine Learning*, **10**(3–4), 142–336.

Jung, H., Pyun, B., & Choi, S. (2012). Combustion and Emission Characteristics of Wood Pyrolysis Oil-Butanol Blended Fuels in a Di Diesel Engine. *International Journal of …*, **13**(2), 293–300.

Kennedy, J., & Eberhart, R. C. (1997). Discrete binary version of the particle swarm algorithm. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, **5**, 4104–4108.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35-45. doi: 10.1115/1.3662552

Kalman, R. E. (1960). On the general theory of control systems. Proceedings of the First IFAC Congress, 1(1), 481-492. doi: 10.1016/s1474-6670(17)63434-1

Khare, A., & Rangnekar, S. (2013). A review of particle swarm optimisation and its applications in Solar Photovoltaic system. *Applied Soft Computing Journal*, **13**(5), 2997–3006.

Lee, K. S., Chin, I. S., Lee, H. J., & Lee, J. H. (1999). Model predictive control technique combined with iterative learning for batch processes. *AIChE Journal*, **45**(10), 2175–2187.

Lintner, W., Tschudi, B., & VanGeet, O. (2011). Best Practices Guide for Energy-Efficient Data centre Design. *U.S Department of Energy*, (March), i–24.

Liu, Y., Wen, J., Liu, G., Chen, L., & Li, Y. (2016). Model predictive control for cooling system of data centres with dynamic environment. Applied Energy, 177, 24-33.

Li, W., Song, Y., Wen, J., & Yan, Y. (2020). An optimal control approach for dynamic cooling management in data centres. Applied Energy, 276, 115436. doi: 10.1016/j.apenergy.2020.115436.

Mahor, A., Prasad, V., & Rangnekar, S. (2009). Economic dispatch using particle swarm optimisation: A review. *Renewable and Sustainable Energy*

Reviews, **13**(8), 2134–2141.

Mäkilä, P. M., & Partington, J. R. (2003). *On linear models for nonlinear systems*. *Automatica*, Vol. 39. doi:10.1016/S0005-1098(02)00183-8

Mayne, D. Q., Rawlings, J. B., Rao, C. V., & Scokaert, P. O. M. (2000a). Constrained model predictive control: Stability and optimality. *Automatica*, **36**(6), 789–814.

Mayne, D. Q., Rawlings, J. B., Rao, C. v., & Scokaert, P. O. M. (2000b). Constrained model predictive control: Stability and optimality. *Automatica*, **36**(6), 789–814.

Mostafa, E., & Mostafa Bentour, E. (2015). Munich Personal RePEc Archive A ranking of VAR and structural models in forecasting A ranking of VAR and structural models in forecasting, (61502).

Ma, Y., Yin, X., & Liu, J. (2018). A model predictive control approach for minimising energy consumption in data centers. Energy and Buildings, 174, 1-12.

Nadjahi, C., Louahlia, H., & Lemasson, S. (2018). A review of thermal management and innovative cooling strategies for data centre. *Sustainable Computing: Informatics and Systems*, **19**(March), 14–28.

Niemann, J., Brown, K., & Avelar, V. (2013). Hot-Aisle vs . Cold-Aisle Containment for Data centres, 1–13.

Niemann, John, et al. “Hot-Aisle vs . Cold-Aisle Containment for Data

Centers." *Whitepaper*, vol. 135, 2013, pp. 1–13.

Pang, W., Wang, K. P., Zhou, C. G., & Dong, L. J. (2004). Fuzzy discrete particle swarm optimisation for solving traveling salesman problem. *Proceedings - The Fourth International Conference on Computer and Information Technology (CIT 2004)*, (May), 796–800.

Parsopoulos, K. E., & Vrahatis, M. N. (2000). Optimal portfolio choice under a liability constraint. *Annals of Operations Research*, **97**(1–4), 131–141.

Patel, C. D., Sharma, R., Bash, C. E., & Beitelmal, A. (2002). Thermal considerations in cooling large scale high compute density data centres. *InterSociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, ITHERM*, **2002-Janua**, 767–776.

Perez, Krystian X., et al. *Integrated HVAC Management and Optimal Scheduling of Smart Appliances for Community Peak Load Reduction*. 2019, pp. 1–21.

Perez, K. X., Baldea, M., & Edgar, T. F. (2019). Integrated HVAC Management and Optimal Scheduling of Smart Appliances for Community Peak Load Reduction, 1–21.

Pisaturo, M., Cirrincione, M., & Senatore, A. (2015). Multiple constrained MPC design for automotive dry clutch engagement. *IEEE/ASME Transactions on Mechatronics*, **20**(1), 469–480.

Pöyhönen, S., Ahola, J., Niemelä, M., Hammo, S., & Punnonen, P. (2021). Variable-speed-drive-based method for the cost optimisation of air filter replacement timing. *Energy and Buildings*, **240**. doi:10.1016/j.enbuild.2021.110904

Prabowo, F. H., Lhaksmana, K. M., & Baizal, Z. K. A. (2018). A multi-level genetic algorithm approach for generating efficient travel plans. *2018 6th International Conference on Information and Communication Technology, ICoICT 2018*, 86–91.

Qin, D. (2011). Rise of var modelling approach. *Journal of Economic Surveys*, **25**(1), 156–174.

Qin, S. J., & Badgwell, T. A. (2003). Process Control Dynamic. *Control Engineering Practice*, **11**, 733–764.

R. Danilak. (2020). Data centre Converged Hardware Is Just A Stop On The Road To A Universal Processor. *Forbes Technology Council*.

Rao, C. v., Rawlings, J. B., & Lee, J. H. (2001). Constrained linear state estimation - A moving horizon approach. *Automatica*, **37**(10), 1619–1628.

Rawlings, J. B. (2000). Tutorial Overview of Model Predictive Control. *IEEE Control Systems*, **20**(3), 38–52.

Richalet, J., Rault, A., Testud, J. L., & Papon, J. (1978). Model predictive heuristic control. Applications to industrial processes. *Automatica*, **14**(5), 413–428.

Riofrío, M. C., Caney, N., & Gruss, J. A. (2016). State of the art of efficient pumped two-phase flow cooling technologies. *Applied Thermal Engineering*,

104, 333–343.

Rob, E. (2021). *The forces behind data centre change - DCD*. *Kingston Technology*. Retrieved from https://www.datacenterdynamics.com/en/opinions/forces-behind-data-center-change/

Salman, A., Ahmad, I., & Al-Madani, S. (2002). Particle swarm optimisation for task assignment problem. *Microprocessors and Microsystems*, **26**(8), 363–371.

Salsbury, T., Mhaskar, P., & Qin, S. J. (2013). Predictive control methods to improve energy efficiency and reduce demand in buildings. *Computers and Chemical Engineering*, **51**, 77–85.

Sargent, T. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *Research, CA Sims - New Methods in Business Cycle 1977, Undefined*. Retrieved from http://www.tomsargent.com/research/sargentsimsBusinessCycleModeling.pdf

Schulz, G. (2011). Server Virtualization. *Cloud and Virtual Data Storage Networking*, 233–248.

Schwenzer, M., Ay, M., Bergs, T., & Abel, D. (2021). Review on model predictive control: an engineering perspective. *International Journal of Advanced Manufacturing Technology*, **117**(5–6), 1327–1349.

Silani, E., & Lovera, M. (2005). Magnetic spacecraft attitude control: A survey and some new results. *Control Engineering Practice*, **13**(3), 357–371.

Song, G., Kowalczyk, D., & Leus, R. (2018). The robust machine availability problem–bin packing under uncertainty. *IISE Transactions*, **50**(11), 997–1012.

Stahl, L., & Sullivan, D. (2001). Computational Fluid Dynamics Modeling of High Compute Density Data centres to Assure System Inlet Air Specifications Chandrakant, 1–9.

Steen, D., Stadler, M., Cardoso, G., Groissböck, M., DeForest, N., & Marnay, C. (2015). Modeling of thermal storage systems in MILP distributed energy resource models. *Applied Energy*, **137**, 782–792.

Theodoridis, Theodoros, and Juergen Kraemer. "System Identification:Theory for the User." Prentice Hall Information and System Sciences Series, 1999.

Tadei, R., Perboli, G., & Perfetti, F. (2017). The multi-path Traveling Salesman Problem with stochastic travel costs. *EURO Journal on Transportation and Logistics*, **6**(1), 3–23.

Tomassetti, G., & Cagnina, L. (2013). Particle swarm algorithms to solve engineering problems: A comparison of performance. *Journal of Engineering (United Kingdom)*, **2013**. doi:10.1155/2013/435104

Urbanucci, L. (2018). Limits and potentials of Mixed Integer Linear Programming methods for optimisation of polygeneration energy systems. *Energy Procedia*, **148**(March), 1199–1205.

Verma, A., Dasgupta, G., Nayak, T. K., De, P., & Kothari, R. (2009). Server Workload Analysis for Power Minimisation using Consolidation. *Proceedings of the 2009 Conference on USENIX Annual Technical Conference*, **1505**, 28.

Wang, W., Rivera, D. E., & Kempf, K. G. (2007). Model predictive control strategies for supply chain management in semiconductor manufacturing. *International Journal of Production Economics*, **107**(1), 56–77.

Wang, X., Yao, Y., Wang, X., Lu, K., & Cao, Q. (2012). CARPO: Correlation-aware power optimisation in data centre networks. In *Proceedings IEEE INFOCOM*, Orlando, FL, USA: IEEE, pp. 1125–1133.

Wold, H. (1964). Econometric model building : essays on the causal chain approach / edited by Herman O. A. Wold | National Library of Australia. Retrieved May 10, 2021, from https://catalogue.nla.gov.au/Record/4233002

XI, Y.-G., LI, D.-W., & LIN, S. (2013). Model Predictive Control — Status and Challenges. *Acta Automatica Sinica*, **39**(3), 222–236.

Vijay K. Vaishnavi, William Kuechler. *Design Science Research Methods and Patterns, 2nd Edition*. 2nd Editio, 2015.

Yang, Z., Qiu, M., Shi, K., & Wu, A. (2019). A quantum-behaved particle swarm optimisation algorithm with extended elitist breeding. *9th International Conference on Information Science and Technology, ICIST 2019*, 496–501.

Yu, N., Salakij, S., Chavez, R., Paolucci, S., Sen, M., & Antsaklis, P. (2017). Model-based predictive control for building energy management: Part II – Experimental validations. *Energy and Buildings*, **146**, 19–26.

Zhang, G., Zhu, X., Bao, W., Yan, H., & Tan, D. (2018). Local Storage-Based Consolidation with Resource Demand Prediction and Live Migration in Clouds. *IEEE Access*, **6**, 26854–26865.

Zhang, X., Lindberg, T., Xiong, N., Vyatkin, V., & Mousavi, A. (2017). Cooling Energy Consumption Investigation of Data centre IT Room with Vertical Placed Server. *Energy Procedia*, **105**, 2047–2052.

Zhao, J., Huang, Z., Sun, L., & He, J. (2020). A hybrid machine learning approach for cooling system optimisation in data centres. IEEE Access, 8, 117261-117272.

Mittal, R., & Iaccarino, G. (2005). Immersed boundary methods. Annual Review of Fluid Mechanics, 37(1), 239-261. https://doi.org/10.1146/annurev.fluid.37.061903.175743

Wu, X., Li, S., Cheng, Z., & Hu, W. (2019). Numerical simulation of airflow and temperature distribution in a data centre using different turbulence models. Energy Procedia, 158, 3692-3697. https://doi.org/10.1016/j.egypro.2019.01.569

ASHRAE Data centre Power Equipment Thermal Guidelines and Best Practices, 2011: https://tc0909.ashraetcs.org/documents/DataCenterThermalGuideline2011_Whitepaper.pdf

"Fundamentals of Heat and Mass Transfer", 7th Edition, by Theodore L. Bergman, Adrienne S. Lavine, Frank P. Incropera, and David P. DeWitt (Wiley, 2011).

"Cooling Strategies for Ultra-High Density Racks and Blade Servers", by Sumit Puri and Ravi Madhavan, presented at the 2008 ASHRAE Winter Conference.

Appendix

This appendix includes supplementary materials for the proposed research. The materials presented in this appendix provide additional details on the experiment code for the VAR-PSO-MPC control framework, the collaboration letter, and the result confirmation letter from the research partner. These documents are essential to understanding the methodology, collaboration, and validation process of the research. The experiment code presents a comprehensive overview of the technical implementation of the VAR-PSO-MPC control framework used in the study. The collaboration letter outlines the collaboration agreement between the research team and the research partner, while the result confirmation letter validates the findings of the research from the field experiment. These supplementary materials aim to provide readers with a complete understanding of the research process and results.

Experimental code

The experiment code presented in this appendix provides a detailed overview of the technical implementation of the VAR-PSO-MPC control framework used in the research study. The code includes all the necessary steps involved in implementing the framework, such as data pre-processing, model training, and optimisation. The code has been carefully commented on and structured to enhance comprehensibility and facilitate reproducibility. The provision of the experiment code in this appendix aims to contribute to the efforts of the scientific community towards promoting open and transparent research practices and facilitates the adoption of the VAR-PSO-MPC control framework in practical applications, thereby promoting the advancement of the field.

Experimental Code

load packages

```
library(dplyr)
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(tseries)
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
library(writexl)
library(openxlsx)
library(forecast)
library(vars)
## Loading required package: MASS
##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##     select
## Loading required package: strucchange
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Loading required package: sandwich
## Loading required package: urca
## Loading required package: lmtest
library(control)
## Warning: package 'control' was built under R version 4.2.3
##
## Attaching package: 'control'
## The following object is masked from 'package:stats':
##
##     step
## The following object is masked from 'package:base':
##
##     append
library(nloptr)
```

Appendix

```
## Warning: package 'nloptr' was built under R version 4.2.3
library(psoptim)
```

Insert initial data

```
#'Data
library(readxl)
data <- read_excel("C:/Users/89776/Desktop/NavyDC_full/initial_data.xlsx",
     col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
         "numeric", "numeric"))
```

insert latest sensors data

```
# set inputs for new iteration
#Define empty matrix for SPs

latest_data <- read_excel("C:/Users/89776/Desktop/NavyDC_full/latest_data.xlsx",
    col_types = c("numeric", "numeric", "numeric",
        "numeric", "numeric", "numeric",
         "numeric", "numeric"))
```

ARIMA models for predicting CPU and outdoor temperature

```
# ARIMA model for outside temperature

arima_t_outside <- function(data) {
  library(forecast)
  l_outside_temp <- log(data[, 4])
  arima_model <- auto.arima(ts(l_outside_temp, frequency = 24), stepwise = FALSE)
  forecast_values <- forecast(arima_model, h = 24)
  T_outside_fcst_matrix<-as.matrix(exp(forecast_values[["mean"]]), ncol = 1, dimnames =    list(NULL, "T-Outdoor"))
 return(T_outside_fcst_matrix)
}

# ARIMA model for CPU temperature
arima_cpu <- function(data) {
  library(forecast)
  l_cpu <- log(data[, 3])
  arima_model <- auto.arima(ts(l_cpu, frequency = 24), stepwise = FALSE)
  forecast_values <-forecast(arima_model, h = 24)
  CPU_fcst_matrix<-as.matrix(exp(forecast_values[["mean"]]), ncol = 1, dimnames = list(NULL, "CPU"))
  return(CPU_fcst_matrix)
}
```

VAR Function to predict PUE and evaluate error.

```
 # Prepare Old data for VAR
 var_data <- as.matrix(data)
  endo_data <- var_data[, 5:8]
  exo_data <- var_data[, 1:4]
  l_endo <- log(endo_data)
  l_exo <- log(exo_data)
```

```
   # prepare new_data for new VAR
  new_data<-rbind(var_data,latest_data)
    new_var_data <- as.matrix(tail(new_data, 167))
    new_endo_data <- new_var_data[, 5:8]
    new_exo_data <- new_var_data[, 1:4]
    new_l_endo <- log(new_endo_data)
    new_l_exo <- log(new_exo_data)

var_model <- function(data, latest_data) {
  library(vars)

  var_model <- VAR(l_endo, type = "const", season = 24, exogen = l_exo,
                  ic = c("AIC", "HQ", "SC", "FPE"))  #' VAR model

  var_pred_endo <- log(latest_data[, 5:8])
  sp_values<-latest_data[,1:2]
  sp_future <- matrix(as.numeric(sp_values), nrow = 24, ncol = 2, byrow
 = TRUE)
  colnames(sp_future) <- c("SP1", "SP2")
  var_pred_exo <-as.data.frame(cbind((sp_future), arima_cpu(data), arim
a_t_outside(data)))
#Convert all list columns to numeric columns
 var_pred_exo[] <- lapply(var_pred_exo, unlist)
 # Apply logarithm to each column except column names
 var_pred_exo[] <- apply(var_pred_exo[], 2, log)

  # initail forecast VAR
     #' Get the names of the exogenous variables in the VAR model
exogen_names <- names(var_model$model$y_exo)

     #' Rename the columns of var_pred_exo to match the exogenous varia
ble names
colnames(var_pred_exo) <- exogen_names

    #' Predict with the modified dumvar
forecast_values <- predict(var_model, newdata = var_pred_endo, dumvar =
 var_pred_exo, n.ahead = 24, ci = 0.95)
 pue_matrix <- forecast_values[["fcst"]][[which(names(forecast_values
[["fcst"]]) == "PUE")]]
pue_fcst <- as.matrix(exp(pue_matrix[,1]))
colnames(pue_fcst) <- "PUE"


  # Check previous forecast against latest actual sensor data
  latest_pue <- exp(latest_data$PUE)
  prev_forecast <- pue_fcst[1]
  current_rmse <- sqrt(mean((latest_pue - prev_forecast)^2))

  # Check if RMSE is greater than 0.4 and update VAR model if necessary
  if (current_rmse > 4) {
   #New VAR model
    new_var_model <- VAR(new_l_endo, type = "const", season = 24, exoge
n = new_l_exo,
                      ic = c("AIC", "HQ", "SC", "FPE"))
    new_var_pred_endo <- log(latest_data[, 5:8])
```

```
    new_var_pred_exo <- log(cbind(sp_future,arima_cpu(new_var_data), ar
ima_t_outside(new_var_data)))

    #Get the names of the exogenous variables in the VAR model
            exogen_names <- names(new_var_model$model$y_exo)

    # Rename the columns of var_pred_exo to match the exogenous variabl
e names
colnames(new_var_pred_exo) <- exogen_names

    #Forecast new VAR
    new_forecast_values <- predict(new_var_model, newdata = new_var_pre
d_endo,dumvar = new_var_pred_exo, n.ahead = 24, ci = 0.95)
    # Forecast PUE value
   new_pue_matrix <- new_forecast_values[["fcst"]][[which(names(new_for
ecast_values[["fcst"]]) == "PUE")]]
new_pue_fcst <- as.matrix(exp(new_pue_matrix[,1]))
colnames(new_pue_fcst) <- "PUE"

# Check if the new model actually improves the RMSE
    new_rmse <- sqrt(mean((latest_pue - new_pue_fcst[1])^2))
    if (new_rmse < current_rmse) {
      # Use the new model for prediction
      var_model <- new_var_model
      forecast_values <- new_forecast_values
      pue_fcst <- new_pue_fcst
    }
  }

  return(list("forecast_values" = forecast_values, "pue_fcst" =pue_fcs
t))
  return(New_data)
}
```

Constraints

```
# Define constraints function
#constraints function
constraints <- function(sp_values, r_temp_min = 18, r_temp_max = 27, pu
e_min = 1.5, pue_max = 2.5 , HG_min = 15, HG_max = 32) {
  forecast_values <- var_model(data, latest_data)$forecast_values
  r_temp <- forecast_values$T_Ambient
  pue <- forecast_values$PUE
  HG<-forecast_values$HG
  g1 <- r_temp >= r_temp_min
  g2 <- r_temp <= r_temp_max
  g3 <- pue <= pue_max
  g4 <- pue >= pue_min
  g5<- HG <= HG_max
  g6<- HG <= HG_min
  c(g1, g2, g3, g4, g5, g6)
}
#Higher penalty constraints
#constraints function
constraints <- function(sp_values, r_temp_min = 18, r_temp_max = 27, pu
e_min = 1.5, pue_max = 2.5 , HG_min = 15, HG_max = 32, HG_penalty_facto
```

```
r = 9) {
  forecast_values <- var_model(data, latest_data)$forecast_values
  r_temp <- forecast_values$T_Ambient
  pue <- forecast_values$PUE
  HG<-forecast_values$HG
  g1 <- r_temp >= r_temp_min
  g2 <- r_temp <= r_temp_max
  g3 <- pue <= pue_max
  g4 <- pue >= pue_min
  g5<- HG <= HG_max
  g6<- HG <= HG_min

  # add higher penalty weight to HG constraints
  constraints_violations <- sum(g1, g2, g3, g4, g5*HG_penalty_factor, g
6*HG_penalty_factor)#'add a higher penalty value for violations of the
HG_min and HG_max constraints.

  return(constraints_violations)
}
#Fitness function
  # create an empty vector to store fitness values
  fitness <- function(sp_values) {
  # calculate the fitness value for the current particle
  constraints_violations <- sum(constraints(sp_values, r_temp_min = 18,
 r_temp_max = 27, pue_min = 1.5, pue_max = 2.5))
  if (constraints_violations > 0) {
    fitness_value <- 100
  } else {
    forecast_values <- var_model(data, latest_data)$pue_fcst
    fitness_value <- sum((forecast_values - 1.5)^2)
  }
  return(fitness_value)
}
```

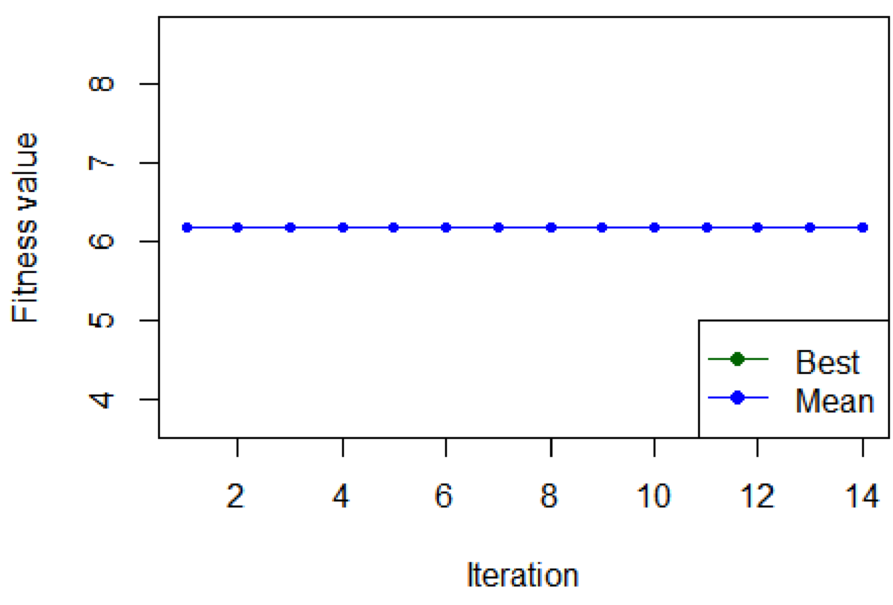
PSO solver

```
library(psoptim)
# Define PSO parameters
n_particles <- 14
n_iterations <- 14
w <- 0.729   # inertia weight
c1 <- 1.494  # cognitive weight
c2 <- 1.494  # social weight
xmin <- c(16, 16) # update the minimum bounds
xmax <- c(28, 28) # update the maximum bounds

# create a logical vector to indicate which dimensions to constrain
vmax_constrained <- c(TRUE, TRUE,TRUE,TRUE, TRUE)

# set the maximum velocity for the constrained dimensions
vmax_values <- rep(5, 2)

# pass the vmax_values to the psoptim function
result <- psoptim::psoptim(FUN=fitness, n=n_particles, max.loop=n_itera
tions, w=w, c1=c1, c2=c2, xmin=xmin, xmax=xmax, vmax=vmax_values, seed=
473, anim=FALSE)
```



#Result

```
# Extract the optimal setpoint values and the minimized fitness value
SP_optimal <- result$sol

fitness_optimal <- result$val

SP_optimal
##              x1       x2
## [1,] 25.20618 23.62396
fitness_optimal
## [1] 6.177956
```

Apply control signal to HVAC system

...

Record results

...

Wait for one hour before running loop again

Sys.sleep(3600) }

Collaboration letter

The collaboration letter included in this appendix outlines the importance of data confidentiality, intellectual property, and ethical considerations in the research partnership between the two parties. The letter aims to demonstrate the researchers' commitment to ethical and responsible research practices and establish a foundation for a successful collaboration. By outlining the roles, responsibilities, objectives, and expected outcomes of the partnership, the letter promotes transparency in the research process. The collaboration letter also ensures that the research process adheres to ethical and legal guidelines, thereby fostering trust and confidence between the research partner and the scientific community.

ที่ กห ๐๕๒๔.๓/๗๘๓

สำนักงานการบริหารเขตพัฒนาพิเศษภาคตะวันออกของกองทัพเรือ (กพอ.ทร.)
(อาคารกองบัญชาการกองทัพเรือ) ถนนอิสรภาพ
แขวงบ้านช่างหล่อ เขตบางกอกน้อย กรุงเทพฯ

๑ มิถุนายา ๒๕๖๕

เรื่อง ขอความอนุเคราะห์ข้อมูลระบบประหยัดพลังงานของศูนย์ข้อมูล(Data Centre)และสร้างความร่วมมือระหว่างหน่วยงานรัฐ

เรียน ผู้บริหารสำนักงานพัฒนารัฐบาลดิจิทัล

ด้วยการท่าอากาศยานอู่ตะเภามีแผนการขยายงานเพื่อเป็นศูนย์การซ่อมอากาศยานขนาดใหญ่แห่งภูมิภาคเอเชีย และรองรับนโยบายตามแผนพัฒนาระเบียงเศษฐกิจภาคตะวันออก (EEC) ของภาครัฐ ซึ่งมีความต้องการสร้างศูนย์ข้อมูลเพื่อรองรับกิจกรรมต่างๆ ที่จะเกิดขึ้นในระบบออนไลน์ และด้วย นาวาเอก ตะวัน บุญคงวัฒนา ตำแหน่งหัวหน้าแผนกแผนการช่าง กองการบินทหารเรือ เป็นบุคลากรที่มีองค์ความรู้ด้านการจัดการพลังงาน ทาง กพอ.เห็นว่าโครงการประหยัดพลังงานของศูนย์ข้อมูลจะเป็นประโยชน์อย่างยิ่งทั้งภาครัฐและเอกชน และทราบว่าหน่วยงานของท่านมีศูนย์ข้อมูลที่ทันสมัยซึ่งควบคุมด้วยระบบอัจฉริยะ

ในการนี้ กพอ.ทร. จึงขอรับความอนุเคราะห์ข้อมูลระบบประหยัดพลังงานของศูนย์ข้อมูล(Data Centre) เพื่อหาแนวทางการสร้างความร่วมมืออันจะก่อให้เกิดประโยชน์ร่วมกันเพื่อเป็นแนวทางขับเคลื่อนประเทศต่อไป

จึงเรียนมาเพื่อพิจารณาเข้าร่วม และขอขอบคุณมา ณ โอกาสนี้

ขอแสดงความนับถือ

พลเรือเอก

(กฤชพล เรียงเล็กจำนงค์)

ผู้ทรงคุณวุฒิพิเศษกองทัพเรือ

รองประธานการบริหารเขตพัฒนาพิเศษภาคตะวันออกของกองทัพเรือ

ผู้ประสานงาน นาวาโท ตะวัน บุญคงวัฒนา
โทร. ๐ ๘๒๘๙๔ ๖๖๔๗
โทรสาร ๐ ๒๔๖๖ ๑๑๘๐ ต่อ ๗๘๒๔๓

Result confirmation letter

The result confirmation letter included in this appendix validates the research findings and confirms their accuracy and validity. The letter highlights the methods used to validate the research results, including statistical analysis, and confirms that the results are reproducible. The confirmation letter provides independent verification of the research findings and serves as a testament to the scientific rigour of the study.

Appendix

No. 0515.11.1/2076

Naval Communication and Information
Technology Department,
Royal Thai Navy, Bangkok, 10600,
Thailand
(066)2475-7700
saraban_mod0507@navy.mi.th

21 October B.E.2565(2022)

Dear Prof Rodgers,

I am pleased to write this corroborating letter regarding the impact that Professor Habin Lee and Ms Yang Hai has had on Royal Thai Navy. Established in 1906, Royal Thai Navy is the naval warfare force of Thailand.

From 2021 to 2022, we worked with Professor Habin Lee and Ms Yang Hai on the EU funded GREENDC project (project number 734273) for energy efficient operations of cooling devices in our data centre. They developed a Model Predictive Control (MPC) framework that predicts future workloads of the servers and generated an optimized operation schedule for Air Conditioners (ACs) in the data centre rooms.

With their support we were able to understand the interaction between heat generated by servers and cold airs provided by ACs. In particular, the framework helped us to change set points of ACs according to the workloads of the servers.

This is particularly important because increasing set point of ACs by 1 degree can result in 4-5% of energy savings. However, we also need to ensure that servers are operating in acceptable cooling conditions. The MPC framework provided us with operational strategies for ACs ensuring appropriate cooling conditions for the servers.

This has directly resulted in:

- 32.5% saving of energy cost. The Royal Thai Navy headquarters' Data Centre used to have the annual cost of electricity approximately 550,000 THB (132,000 GBP). The research by GREENDC illustrates an annually energy saving potential of 178,416 kWh, this will result in a cost saving approximately 900,000 THB (21,600 GBP). It is a very positive and encouraging results, and we are expecting to cooperate with your research team in the future. The saved electricity budget can be used to invest to upgrade the hardware in our data Centre. Also, we expect the intelligent energy saving control solution to be released and put in use.

- Increased energy awareness of DC staff members. Through the experiments with the Brunel team, the staff members of the DC were able to integrate energy efficiency in their business processes. They also collaborate with other DCs to increase energy efficiency based on the outcome of the project.

We have no doubt that without Professor Lee and Ms Hai's contributions, we would not be able to achieve the energy savings for the operation of our data centre.

Yours Sincerely,

Vice Admiral

(Chayut Navesaphutikorn)

Director of Naval Communication and Information

Technology Department