Energy Optimisation in smart Data Centers using Machine Learning Techniques

Jyotsna Gaur
Dept of Science and Technology
University of Canberra
Canberra, Australia
u3248720@uni.canberra.edu.au

***Abstract*—One of the most complex challenges of running and managing a Data Center is optimising the energy needed to run the facility. Each of the Data Center components such as Power subsystems, Uninterruptible power supplies (UPS), Backup generators, Ventilation and cooling equipment, Fire suppression systems and Building security systems consume a large amount of energy. Many aspects of Data Center technology require high amounts of energy consumption and this needs efforts in order to be managed efficiently. This paper describes design science research and a conceptual framework for building a machine learning model to predict and optimise energy consumption in a Data Center facility. The model will be based on developing a Convolutional Neural Network (CNN) model to determine the power usage effectiveness (PUE) of a data center infrastructure. The model could be further validated against energy generation data for the past year to get the accuracy matrix and descriptive statistics for the model efficiency and obtain the Power Usage Effectiveness (PUE) of the Data Center. The paper also explores computational fluid dynamics and its simulation which can be used in gathering the micro-climatic data for the Data Center and testing various cooling techniques to develop prototypes for field evaluation.**

Keywords—Energy optimisation in Data Center, Machine Learning in Data Center, Power Usage Effectiveness (PUE), Energy Load, Green IT, Power efficiency, Data Center Cooling Simulation.

I. INTRODUCTION

A Data Center is a physical location that stores computing machines and their related hardware equipment. It contains the computing infrastructure that IT (Information Technology) systems require, such as servers, data storage drives, and network equipment. It is the physical facility that stores any company's digital data. [1] Previously every company invested in and managed their own data center facility. Over the years, the size and power requirements of computers has reduced. But the demand for these services-e.g., computing, storage, networking with minimum latency; has risen exponentially. This has led to an increase in the rapid expansion of Data Center infrastructure and operational costs as well as energy consumption. Data centres contribute around 0.3 per cent to overall carbon emissions, while the ICT sector accounts for more than 2 per cent of global emissions. [2] This percentage is growing exponentially as the years go by. Rising energy costs and environmental responsibility has put tremendous pressure on the Data Center industry to improve its operational efficiency.

This paper gently introduces the planned research design ideas for optimisation of energy usage in Data Centers. This research will lead to a better understanding and management of Data Centers. It will help bring down the running energy costs of their cooling infrastructure. Investing in this research will lead to better optimized Data Centers. A network engineer could take care of a futuristic Data Center by simply monitoring it through the cloud and pinpoint the exact issues and places of concern while maintaining it on-site or off-site. The stakeholders can be localized Data Center companies which function locally in Australia or the global technology players like Google and Amazon which have their Data Centers spread across various regions in the world. The infographic below describes certain characteristics of a modern datacenter.

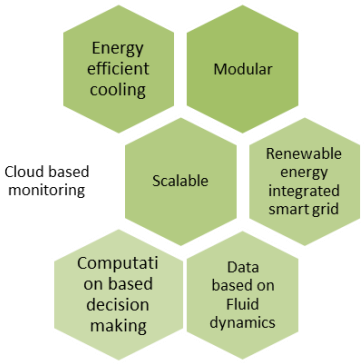

Figure 1: Smart Data Center features, source: Author

II. LITERATURE REVIEW OF ENERGY EFFICIENCY INCREMENTING OPTIMIZATION METHODS

A. *Thermal optimisation*

Thermal optimization of the data centers is important to maintain the temperature, which leads to optimum energy conservation. In their paper, more than 10 data centers have been studied by Ni and Bai [1] by considering the energy consumption of the cooling systems. It has been observed by them that more than 50% of the data centers were not operating efficiently, leading to energy wastage.

This wastage can be mitigated by developing passive and active strategies like humidity control, temperature control, optimisation of airflow and economic cycles. Some of these energy saving techniques have been compared by Oro et al [4]. They found that integrating the energy grid with recovery of waste heat through an absorption-refrigeration system or through direct power generation and indirect power generation made a huge difference in energy usage.

Some passive cooling techniques for data centers have been studied by Daraghmeh and Wang [2]. These techniques included applications of the air, water and heat pipe which were incorporated with other operations like solar systems, absorptions, evaporative cooling, and geothermal cooling.

B. Machine Learning and AI based optimisation

Multiple technologies like network optimizations, servers and processors have been proposed to improve the efficiency of the data centers. The usage of all these technologies can generate historical databases that can later be analyzed and modelled for making future predictions.

The most notable work is where a neural network methodology has been proposed by Gao [5] for energy optimization of the data centers. As a result of multiple configurations and non-linearity, it is difficult to optimise the energy efficiency of the data centers, which is where a neural network seems to be the most useful model. The model gets trained from the actual operational data for modelling the performance of the data center and to predict the Power Usage Effectiveness (PUE) accurately. A PUE of 1.1 has been obtained with an error of 0.4%. This work has been tested and validated in the Google Data centers and it has been seen that AI techniques are effective ways to optimise the performance and improve the energy efficiency.

The table below summarizes some of the machine learning techniques that have been researched in the past for Data Centres:

TABLE 1: COMPARISON OF VARIOUS ML ALGORITHMS FOR DATA CENTERS, SOURCE: KUMAR ET AL [3]

S. No.	Type of Algorithm used	Comments	Author
1	ANN	Power Usage effectiveness is calculated. It is validated in Google Data centers	Gao [5]
2	Naïve Bayes	Implemented on rack mounted servers on data center testbed	Marco et al [6]
3	Multivariable Linear Regression	FIESTA IoT and Resl DC testbed were used for the implementation. ANN and SVM were considered as a future scope	Smpokos et al [7]
4	Reinforcement learning	Airflow regulation was performed. It has a high computation time.	Lazic et al [10]
5	ANN	The effect of the cooling parameters was not considered.	Yu et al [9]
6	Deep Deterministic Policy Gradient	It has been used for both solving simple optimisation and complicated optimisations like the data center cooling	Lillicarp et al [8]

The above machine learning techniques and their selection vary on a case-by-case basis. But there is a strong presence of neural networks while using data that is organic in nature and this is true for energy output data. The variables can be based on different micro-climatic factors in a data center. These can be the ambient temperature, humidity, water consumption, fan speeds, radiation, etc. Analysing the microclimate of a data center premise and estimating its cooling effectiveness is not a novel but definitely a niche aspect of research related to energy optimisation. The proposed study will analyse the climatic factors and different energy outputs to evaluate the most effective variables contributing to cooling of data centers. We can then control these variables to come up with efficient models that can correctly predict the PUE in a data center.

III. PLANNED RESEARCH DETAILS

A. Planned Research Design

The aim of this planned research is to study and create a research framework for energy optimisation in smart Data Centers using Machine Learning models. It involves the objectives such as- identification of how computational fluid dynamics can be used for simulation of ventilation in Data Center. It involves understanding the various variables involved in creating an energy efficient Data Center. After a suitable literature review and understanding of core concepts, a suitable machine learning model can be created for predicting the PUE of a Data Center.

The research questions are as follows:

i. How can computational fluid dynamics and its simulation be used in gathering the micro-climatic data for the Data Center?

ii. How to design and train an appropriate energy model to predict the cooling efficiency of a Data Center through artificial intelligence?

iii. Can we predict the most energy efficient modular arrangement for a Data Center covering a defined area?

The proposed research aims to provide a conceptual framework, where the author talks about the argumentation for research, explanation of issues and generation of solutions. Apart from the argumentation stated in above sections, there are several issues that can be addressed by this research. The issues have been identified by reading the available literature on the topic and identifying research gaps. There are issues pertaining to selecting the most optimised cooling infrastructure for Data Centers, issues related to monitoring of these cooling infrastructures and issues related to scheduling of server loads and server consolidation. The solutions will be prediction of most optimum states of cooling infrastructures using neural networks on existing micro climatic data of Data Centers.

Through this research, a framework proposing an energy optimisation policy based on machine learning will be introduced to predict the Power Usage Effectiveness (PUE) of Data Centers. This would help the stakeholders to anticipate the energy requirements for cooling and the power consumption before placing or moving jobs, and therefore choose a job configuration that is expected to be good. This power of decision making can prove useful for the Data Center operating industry as well.

The research will take a quantitative approach towards research design. The study will have a quantitative approach because it is mostly dealing with quantitative variables and the data is mostly numeric. In the quantitative approach, there will be a machine learning model based on neural networks that will be built. The model will have input variables pertaining to micro-climatic factors of Data Center premises. The research method will be based on Model Building using machine learning.

B. Planned Sample

The data sample for research design will be derived on site at the Data Center facility. For this appropriate permissions and consent will be taken from the Data center operators. Before collecting the data, the data requirements shall be

framed. These requirements include the different types of micro-climatic data at the Data Center. The micro-climatic data can be gathered by attaching sensors at critical locations within the Data Center infrastructure. These sensors can capture the ambient temperature, humidity, wind speed, cooling load, building energy consumption in units, etc. These data will have quantitative variables. This data can also be used to validate the machine learning model which would be created from the simulation data.

The data requirements also include the dimensions of the Data Center building, along with rack heights/dimensions and server room dimensions and climatic data. These dimensions will be used to create a 3d Building model which could be used for creating simulations using Building and climate analysis softwares. These simulations can include building energy usage estimated values in different seasons and under different loads. It can also give indications towards creating a simulated cooling technique and its alternatives.

All these building simulations will generate another set of data which can be used for predictive modelling using machine learning. This data will help to create and train the desired model for predicting the energy load on the Data Center.

C. Planned Instruments for Data Collection

The primary sources for data collection involve experimental observations based on building simulations. The instruments will be the observations which are derived from sensors attached to the Data Center facility.

TABLE 2: DESIGN SCIENCE CHART FOR PROPOSED RESEARCH, SOURCE:AUTHOR

Stage	Goal	Output	Methods
1.Problem Identification	Identify the cooling challenges faced by data centers	Problem statement	Literature review, Expert interviews, Site visits
2.Conceptual design	Develop a conceptual framework for efficient cooling	Conceptual model	Brainstorming, Concept mapping, Systematic review, Machine Learning model
3.Prototype development	Develop a cooling system prototype using Building simulation and 3d modelling	Working prototype	Experimental design, Simulation, Testing
4.Evaluation and refinement	Evaluate and refine the cooling system prototype	Refined prototype	Performance testing, User feedback, Energy consumption analysis
5. Implementation	Implement the cooling solution in a test data center	Implemented solution	Deployment, Monitoring, Documentation
6. Report	Communicate the research findings and practical implications	Research report, Presentation, Publication	Research report, Presentation, Publication

The secondary sources of data collection will involve relevant documentation, videos and literature based on Data Centers. The instruments used for this type of data involve reading case studies- such as written documents about energy optimisation in Google Data Centers. There will be other secondary sources such as research papers and interviews with industry experts which will also help in gathering concepts and knowledge on Data Center technology.

The table 2 explains the design science chart to be employed in creating a model for determining the power usage effectiveness of a Data Center.

D. Planned Variables in the Study

The nature of the data collected will be quantitative observation-based data. Mostly all the variables derived from the micro-climatic data will be possibly gathered and used in the study as follows, source [5]:

i. Total server IT load [kW]: The total amount of power consumed by all the servers in the data center.
ii. Total Campus Core Network Room (CCNR) IT load [kW]: The total amount of power consumed by the networking equipment located in the data center's core network room.
iii. Total number of process water pumps (PWP) running: The number of pumps used to circulate the water used in the cooling process.
iv. Mean PWP variable frequency drive (VFD) speed [%]: The average speed of the PWP's variable frequency drive, which regulates the flow of water in the cooling system.
v. Total number of condenser water pumps (CWP) running: The number of pumps used to circulate water through the condenser in the cooling system.
vi. Mean CWP variable frequency drive (VFD) speed [%]: The average speed of the CWP's variable frequency drive, which regulates the flow of water through the condenser.
vii. Total number of cooling towers running: The number of cooling towers used in the cooling system.
viii. Mean cooling tower leaving water temperature (LWT) setpoint [F]: The setpoint temperature of the water leaving the cooling towers.
ix. Total number of chillers running: The number of chillers used in the cooling system.
x. Total number of dry coolers running: The number of dry coolers used in the cooling system.
xi. Total number of chilled water injection pumps running: The number of pumps used to inject chilled water into the cooling system.
xii. Mean chilled water injection pump setpoint temperature [F]: The setpoint temperature of the chilled water injected into the cooling system.
xiii. Mean heat exchanger approach temperature [F]: The difference between the chilled water supply temperature and the temperature of the air entering the heat exchanger.
xiv. Outside air wet bulb (WB) temperature [F]: The temperature of the air when measured using a wet-bulb thermometer, which takes into account the humidity of the air.

xv. Outside air dry bulb (DB) temperature [F]: The temperature of the air when measured using a standard thermometer.

xvi. Outside air enthalpy [kJ/kg]: The total energy content of the outside air per unit of mass.

xvii. Outside air relative humidity (RH) [%]: The amount of moisture in the outside air relative to the maximum amount of moisture that the air could hold at that temperature.

xviii. Outdoor wind speed [mph]: The speed of the wind outside the data center.

xix. Outdoor wind direction [deg]: The direction from which the wind is blowing outside the data center.

E. Plan for Data Analysis

The data collected from this research would be analysed according to the research pipeline shown in figure 2.

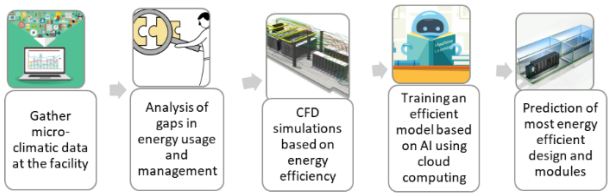

Figure 2: Proposed Research and Data analysis pipeline, source:Author

Data obtained from the real time data collection using sensors at the data center will be prepared for further analysis. All the observations regarding the micro climatic variables at the data center will be checked for outliers, missing values. Also, we are obtaining 2 sets of data in our research. The first set will be the real time microclimatic data which would be used for model evaluation. The second is the data generated after simulations which will be used to create and train the machine learning model to predict the power usage effectiveness (PUE) of the data center.

After preparing the data, statistical analysis and data visualization will be done to further learn about the nature of the data and determine the best type of machine learning model to be used. The nature of data can also be estimated after plotting the data. It can be either organic or follow a curved or specific pattern. An analysis of the gaps in energy usage and management through graphs and trendlines can also be executed to understand the data better.

The cleaned and prepared data will then be used to evaluate a machine learning model created with the simulation data as described in the following paragraph.

A separate Building Simulation dataset will yield simulation data with the same variables. This dataset can be used to create and train a machine learning model to predict the power usage effectiveness under different circumstances such as applying different cooling techniques and checking the variable values under different seasons of the year. This model would then be used to analyse and predict the power usage effectiveness of the Data Center.

F. Plan for Data Presentation

There would be 2 types of data presentations that could be done in the research. First is presenting the collected data from the sensors and building simulations. This can be done by the use of drawing relationship graphs between different climatic variables of the Data Center. The use of scatterplots can also be implemented to study and analyse the data for different patterns and deriving equations which could result in determining the modelling techniques to be used.

The second type of data presentation includes displaying of model validation results and accuracy matrix. This could be done with the use of graphs and tables. The research could also yield in a poster presentation which can be done to display the research ideas to a conference or seminar.

G. Evaluation/Validation strategies used

The strategy for validating research on data center will involve experimental validation of the trained machine learning model. This involves designing and carrying out experiments that replicate the conditions and parameters of the data center cooling system to validate the findings. The dataset used for validation will be the realtime climatic data derived from the sensors attached to data center infrastructure.

The most appropriate cooling techniques could also be replicated in the physical environment after deriving results from the thermal simulations done using software. These could also be tested and measured to further validate the best cooling alternatives. Hence a field-based validation could also work in this case.

The trained model from the machine learning algorithm will also have to be validated using techniques such as cross validation and k-fold validation. In this case 80 percent of the data can be used for training and 20 percent can be used for testing. The number of folds for cross validation may vary from 1,3,5 or more as per need.

H. Strategies used and the proof-of-concept evaluated for effective modular design

Another aspect of this research is to develop the most efficient modular design for a Data Center. For this purpose, a proof-of concept using building simulation can be made for different types of modular design of a Data Center. It is always convenient to create a proof of concept first before carrying out a field validation.

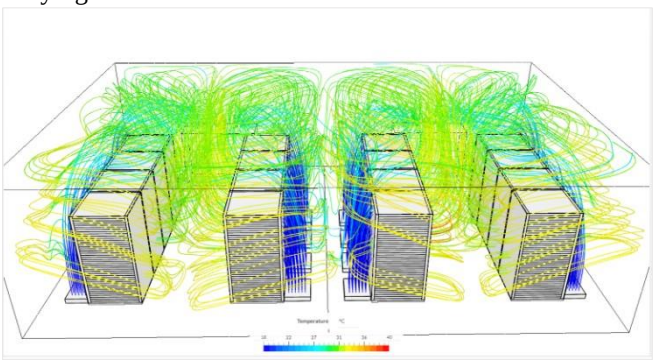

Figure 3: Data Center thermal simulation, source: simscale.com

The following strategies may be involved for this proof of concept:

i. Defining the specific aspects of the modular design that need to be tested. This could include

testing the scalability, flexibility, and efficiency of the modular design.

ii. Developing a prototype of the modular Data Center design based on the defined scope and objectives. This could involve using modular components such as modular racks, cooling systems, and power distribution units.

iii. Conducting testing and validation of the modular design prototype in a controlled environment such as a test lab or a small data center. This could involve testing the performance of the modular components such as cooling systems, power distribution units, and racks, and validating the design against the defined objectives.

iv. Evaluating the performance of the modular design based on metrics such as power usage effectiveness (PUE), energy efficiency, and scalability.

v. Optimization and improvement could be done to improve the modular design prototype to achieve the defined objectives. This could involve making changes to the cooling system, power distribution, and rack design to improve efficiency, scalability, and flexibility.

vi. Preparing a report which can be used to communicate the results of the POC evaluation to stakeholders and to guide the development of a full-scale modular data center design.

Overall, a POC evaluation can provide valuable insights into the effectiveness of a modular design concept for a data center. It can help to identify design issues, optimize the design, and provide recommendations for further development.

IV. Conclusion and Future Work

The research would lead to better prediction of Power Usage Effectiveness (PUE) of Data Centers. This would help in energy optimization of the cooling infrastructure of Data Centers. The research would contribute to both the theory and practice of designing cooling infrastructures of Data Centers. In terms of theory, a policy framework will be devised tabulating the best practices and construction guidelines which can be further researched on. In terms of practice, the ML model can be run for various data sets of various Data Centers.

Limitations: The scope and the variables need to be clearly assessed and selected. In the above steps there is a provision of possible variables, but they are based on the research on Google Data Centers. The variables will be different based on the Data Center we will get to model. Also, the acquisition of datasets pertaining to Data Centers is highly classified and is difficult to obtain from a company. Possibly sensors can be used which can be part of this research, to record the data on the author's end, since companies generally do not share the information.

A. Ethical Issues

Data Centers store a massive amount of sensitive and confidential information. Energy optimization research could inadvertently access and expose this data. Researchers need to ensure that data security protocols are in place to protect the privacy of the data.

Energy optimization research aims to reduce the amount of energy consumed by Data Centers. However, some methods of reducing energy consumption, such as the use of renewable energy sources, can have unintended environmental impacts. For example, the production and disposal of solar panels can have negative environmental effects. Researchers need to weigh the environmental impact of their methods against the benefits they provide.

Energy optimization research may inadvertently introduce bias into the systems it aims to optimize. For example, if the research is conducted on a dataset that is not diverse, the resulting energy optimization system may not work well for all users. Researchers need to be aware of potential biases and work to eliminate them.

Energy optimization research can benefit large Data Centers with significant financial resources to invest in energy-efficient technologies. However, smaller Data Centers and organizations may not have the same resources, putting them at a disadvantage. Researchers need to consider the impact of their methods on all organizations, including those with fewer resources.

Energy optimization research can have significant impacts on the operation and management of Data Centers. Therefore, it is essential to ensure that the research is transparent and that Data Center operators understand how the methods work. This transparency can help to build trust and ensure that the research is used responsibly.

B. Research Quality

In energy optimization research, it is essential to ensure that the methods used to optimize energy consumption are valid and reliable. For example, if the research measures the effectiveness of a specific technology, it is crucial to ensure that the technology is accurately and appropriately evaluated. It is essential to ensure that the findings are generalizable across different data centers and that the methods used can be replicated in other settings.

It is crucial to ensure that the data collected for the study is representative of the data center population being studied. For example, if the study only focuses on large data centers, the findings may not be applicable to smaller data centers.
Energy optimization research can raise ethical concerns related to data privacy, informed consent, and confidentiality. It is essential to ensure that research is conducted ethically, and that participants' rights are protected. It is essential to use appropriate statistical methods and ensure that the results are interpreted correctly. There is a need to ensure that the methods used can be replicated, and the findings can be validated by other researchers. In energy optimization research, it is essential to undergo a rigorous peer review process to ensure the quality of the research.

Acknowledgment

I would like to acknowledge and thank Prof Dharmendra Sharma and Dr Asmaa Elsaiedy for their constant support and constructive feedback.

References

[1] J. Ni and X. Bai, "A review of air conditioning energy performance in data centers," Renew. Sustain. Energy Rev., vol. 67, pp. 625–640, Jan. 2017.

[2] H. M. Daraghmeh and C.-C. Wang, “A review of current status of free cooling in datacenters,” Appl. Therm. Eng., vol. 114, pp. 1224–1239, Mar. 2017.

[3] Kumar, R., Khatri, S.K. and Diván, M.J., 2020, July. “Effect of cooling systems on the energy efficiency of data centers: Machine learning optimisation” In *2020 International Conference on Computational Performance Evaluation (ComPE)* (pp. 596-600). IEEE.

[4] E. Oró, V. Depoorter, A. Garcia, and J. Salom, “Energy efficiency and renewable energy integration in data centres. Strategies and modelling review,” Renew. Sustain. Energy Rev., vol. 42, pp. 429–445, Feb. 2015.

[5] J. Gao, “Machine Learning Applications for Data Center Optimization,” Google white Pap., pp. 1–13, 2013.

[6] V. S. Marco, Z. Wang, and B. Porter, “Real-Time Power Cycling in Video on Demand Data Centres Using Online Bayesian Prediction,” in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 2125–2130.

[7] G. Smpokos, M. A. Elshatshat, A. Lioumpas, and I. Iliopoulos, “On the Energy Consumption Forecasting of Data Centers Based on Weather Conditions: Remote Sensing and Machine Learning Approach,” 2018.

[8] T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” Mach. Learn., Sep. 2015.

[9] W. Yu, Z. Wang, Y. Xue, L. Guo, and L. Xu, “A combined neural and genetic algorithm model for data center temperature control,” CEUR Workshop Proc., vol. 2252, pp. 58–69, 2018.

[10] N. Lazic et al., “Data center cooling using model-predictive control,” Adv. Neural Inf. Process. Syst., vol. 2018-Decem, no. NeurIPS, pp. 3814–3823, 2018.