Optimization of Power Usage Effectiveness for Heterogenous Modular Data Centers using Neural Network

Vishal Kumar Singh
College of Engineering and Computer Science
University of Michigan-
Dearborn, USA
e-mail: vksingh@umich.edu

Dr. Jinhua Guo
College of Engineering and Computer Science
University of Michigan-
Dearborn, USA
e-mail: jinhua@umich.edu

***Abstract*- With the rise of Internet of Things (IoT), it is becoming cheaper and easier to collect data from data center (DC) mechanical, electrical and control systems. These systems have complex interactions with each other. The static control logics and high number of configuration and nonlinear interdependency create challenges in understanding and optimizing energy efficiency. This is particularly challenging and expensive in medium size or smaller configurations like data suites or modular data centers. We utilize a learning engine that learns from operationally collected data to accurately predict power usage effectiveness (PUE) and create a control model to validate test results. Using the machine learning framework developed in this paper, we are able to predict DC PUE within 0.0004 +/¬ 0.0005. The results show that machine learning can improve data suite efficiency. The results also indicate that neural network based controller shows promise for practical implementation.**

Keywords— Machine learning; Neural Network; PUE; Data center.

I. INTRODUCTION

Data centers are recognized as an increasingly troublesome percentage of electricity consumption in the US. A recent revision of the Koomey report [1] puts this at 2% of all US power consumption and 1.3% of worldwide power consumption. Rapid growth of cloud based systems is accelerating growth of data centers. Growing energy costs and environmental responsibility have placed the DC industry under increasing pressure to improve its operational efficiency. The development of metrics of data center efficiency (e.g., PUE) has focused attention on improving energy efficiency in data centers. Even large companies have scored low on Greenpeace report.

Figure 1. Examples of containerized/modular data center

Constructing data center space using traditional methods takes a long time. Speed of delivery of data center space has become a critical business factor for data center operators. This gave rise to modular data centers and containerized data centers. Figure 1. shows few examples of modular data center. Many companies build modular data center for inside building shell and standalone containers for outside [17] [18] [19].

Aisle containment has improved efficiency of facility side cooling power usage (chiller and fan) and load balancing of virtual servers has improved server power usage consumed by servers (IT Load) [2].

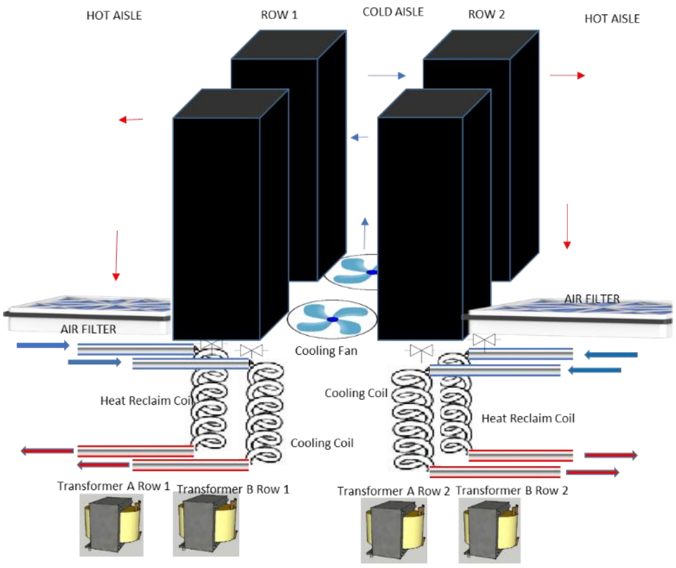

Figure 2. 2-D heterogeneous DC Experimental setup design

Internet of Things (IoT) is rapidly growing with projected \$7.1 trillion by 2020 [3]. This has allowed for significant changes in asset instrumentation to communicate via internet protocol (IP). By using IoT framework, it has become possible to collect and analyze granular data from uninterruptible power supply (UPS), computer room air conditioning (CRAC), circuits, power distribution unit (PDU) etc. This allows collecting data from smaller sections of data centers like aisle, suite or data pod. Instrumenting these microsystems has allowed to manage and control smaller environment ecosystems in a data center. Figure 2. shows a modular data suite architecture and data collection points in a midsize data center.

At the given scale of power use, any incremental improvements in efficiency will produce notable cost savings

 ISBN: 978-1-61208-484-8

and reduce carbon emissions. Large data centers like Google, Microsoft and Amazon have homogenous standard systems as compared to smaller privately held multi-tenant data center that have heterogeneous (non-standard) systems. Aisle containment and efficient virtual server load management have attempted to improve energy efficiency in data centers [3]. Metrics devised more recently like Corporate Average Data Center Efficiency (CADE) have drawn attention more broadly to all power consumption in the data center, including both cooling systems and servers, showing that there is still significant underutilization in data centers [4]. Recently, more efforts are being made to optimize data center efficiency by utilizing machine learning [5].

This paper focuses on using neural network optimizing method to predict and optimize cooling power of a given load in a modular heterogeneous data suite to optimize overall PUE.

In Section II we discuss related work. In Section III, we discuss the methodology of the neural network approach. In Section IV, we discuss results and discussion. In Section V, we discuss limitations of machine learning. Finally, we conclude in Section VI.

II. Related Work

Increasing energy efficiency in a data center has been in great focus in the past few years. Efforts and have been made to optimize facilities by aisle containment [6]. There also has been work on managing virtual server loads to utilize energy efficiently [7]. There is work done in managing energy by combining building automation and virtualization together [2].

There are new demands around cloud computing, big data and infrastructure power efficiency. Furthermore, this change in the data center is being driven by more users, more data and a lot more reliance on the data center itself.

With cloud technologies and the rapid growth in data leading the way within many technological categories — working with the right data center optimization technologies has become more important than ever [9]. Data center Administrators must understand where their current energy demands are allocated and how they can best optimize those resources. Every small amount of energy efficiency gains is improvement. Recently Microsoft and Google have used machine learning techniques for energy optimization. Microsoft is measuring server workload spikes and automating data center operations [8]. Google is exploring using machine learning techniques to optimize energy use data center at a building level [5]. There has been no application of machine learning techniques in a mid-size data centers. This is due to lack of instrumenting machines and implementing IoT platform to collect and store data. The facility side infrastructure has components that have complex interactions amongst themselves. Most of the existing optimization techniques use static method such as cold aisle set point temperature. Establishing an accurate mathematical model or obtaining characteristic parameters for a proportional–integral–derivative (PID) controller in practical control scenarios is challenging, thus limiting their practical applicability [20]. On the other hand, machine learning can be accurately modeled to represent true characteristics of a DC. All the related studies for midsize data centers have been using simulations, we show results by collecting data from practical operations in mid-size data center. This study is unique in applying machine learning energy optimization technique on facility side infrastructure operational data in midsize modular data center.

This study relates to micro systems like data suites and modular data center in multi-tenant facility with heterogeneous server configurations, see Figure 2. This study is to further optimize micro facility environment related to a data suite for a given server load.

III. Methodology- Machine Learning Approach

Facility side infrastructure has components that have complex interactions amongst themselves. PID models do not accurately capture these interactions. Machine learning is well-suited for the DC environment given the complexity of plant operations and the abundance of existing monitoring data. The modern large-scale DC has a wide variety of mechanical and electrical equipment, along with their associated set points and control schemes. The interactions between these systems and various feedback loops make it difficult to accurately predict DC efficiency using traditional engineering formulas. We are training the neural to produce optimal set of operating parameters. Rectified Linear Units (ReLU) is used for deep learning. The model is trained to optimize for lowest PUE.

Neural Network is the machine learning approach which uses Multi-Layer Perceptron (MLP), Supervised Learning and Resilient Back Propagation Algorithm to make an efficient prediction of PUE $P_\theta(x)$ using the environmental variables n that surrounds heterogeneous DC, such as Cold Coil Temperature, Cold Aisle Temperature, Cooling Coil Chilled liquid flow, Fan Power, Chiller Power, Server Load, etc. Let us consider an x as a set of input $m \times n$, where m is the size of the dataset and n is the number of features. The input matrix is then multiplied with the model parameter θ to give the hidden layer. The size and number of hidden layers can be varied based on the complexity of the model required.

The Neural Network is adapted to DC through mathematical model framework for training DC energy efficiency models. Neural networks are a class of machine learning algorithms which adapt and react based on the behavior of neurons. They have best fit adaption, pattern searches and so on to accommodate the accuracy. The concept of machine learning is explained in detail with implementation.

A. Multi Layer Perceptron

The neural network algorithm used multi-layer perceptron, which is well applicable when modeling functional relationships. The underlying structure of an MLP is a directed graph, i.e., it consists of vertices and directed edges, in this context called neurons and synapses [10]. The neurons are organized in layers, which are usually fully connected by synapses. The synapse can only connect to subsequent layers. The input layer consists of all covariates in separate neurons and the output layer consists of the response variables. The layers in between are referred to as hidden layers, as they are not directly observable. Input layer and hidden layers include a constant neuron relating to intercept synapses, i.e. synapses that are not directly influenced by any covariate. Figure 3 gives an example of a neural network with one hidden layer that consists of three hidden neurons. This neural network models the relationship between the two covariates A, B and the response variable Y. Theoretically allows inclusion of arbitrary numbers of covariates and response variables. However, there can occur convergence difficulties using a huge number of both covariates and response variables.

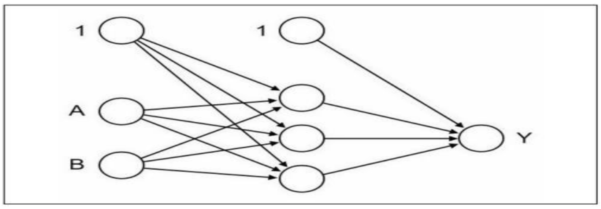

Figure 3. Example of a neural network.

To each of the synapses, a weight is attached indicating the effect of the corresponding neuron, and all data pass the neural network as signals. The signals are processed first by the so-called integration function combining all incoming signals and second by the so-called activation function transforming the output of the neuron.

The simplest multi-layer perceptron (also known as perceptron) consists of an input layer with *n* covariates and an output layer with one output neuron.

It calculates the function

$$o(x) = f(w_o + \sum_{i=1}^{n} w_i x_i) = f w_o + w^T x \qquad (1)$$

where w_o denotes the intercept, **w** = (w_1,..., w_n) the vector consisting of all synaptic weights without the intercept, and **x** = (x_1,..., x_n) the vector of all covariates.

B. Supervised Learning

Neural networks are fitted to the data by learning algorithms during a training process which focuses on supervised learning algorithms [13]. These learning algorithms are characterized by the usage of a given output that is compared to the predicted output and by the adaptation of all parameters according to this comparison. The parameters of a neural network are its weights. All weights are usually initialized with random values drawn from a standard normal distribution.

C. Backpropagation And Resilient Backpropagation

The resilient backpropagation algorithm is based on the traditional backpropagation algorithm that modifies the weights of a neural network in order to find a local minimum of the error function [14].

D. Implementation

The machine learning algorithm used is Neural Network. The neural network utilizes 2 hidden layers and 0.01 as the regularization parameter. The training dataset contains 19 input variables and one output variable (the Suite PUE) as shown in the Figure 5b. The total size of the data samples used is 119421 rows, which were collected from a heterogeneous data center sensor ports. The 70% of the dataset is used for training with the remaining 30% used for cross-validation and testing. The chronological order of the dataset is randomly shuffled before splitting to avoid biasing the training and testing sets on newer or older data [15].

The 19 variables used for modelling are as follows.

TABLE I. SELECTED VARIABLES

Variables	Variables
Cooling Coil Leaving Temperature (°F)	Power utilized by chilled liquid (kW)
Average Cold Aisle Temperature (°F)	Cold Coil IN Water Temp (°F)
Cooling Coil Valve Position (%)	Cold Coil Out Water Temp (°F)
Fan Speed (%)	Server Load A (kVA)
Hot Aisle Temperature (°F)	Server Load A (kVAR)
Heat Reclaim Coil Leaving Temperature (°F)	Server Load A (kW)
Cooling Coil Chilled liquid flow (Gallons/Min.)	Server Load B (kVA)
Absorption (kW)	Server Load B (kVAR)
Fan Power (kW)	Sever Load B (kW)
	Suite Server Load (kW)

Data normalization, also known as feature scaling, is recommended due to the wide range of raw feature values. The values of a feature vector *z* are mapped to the range [-1, 1] by:

$$z_{norm} = \frac{z - mean(z)}{\max(z) - \min(z)} \qquad (2)$$

The Block diagram explains the overall scenario acquired in the Data center for Predicting PUE, based on the Machine Learning Algorithm Neural Network Model.

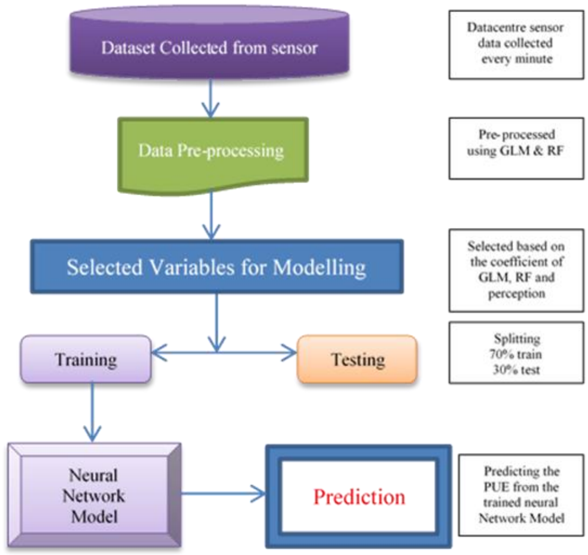

Figure 4. Block diagram of Neural Network Modelling

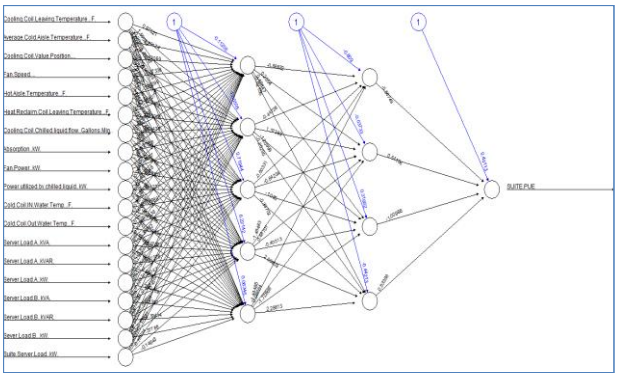

Figure 5. Network Model with selected variables

The block diagram as shown in Figure 4. represents the logic flow of neural network prediction model which evolves the processing of data retrieved from the sensor ports. The data variables are of different features which may or may not affect the SUITE PUE. The collected data is preprocessed through Generalized Linear Model (GLM) [11] and Random Forest (RF) [12] algorithm to find the effectiveness of the parameter with the coefficients. The variables are selected from the preprocessed data through positive skewness arrived with the target SUITE PUE. This achieved through the Generalized Linear Model (GLM), Random Forest (RF) and Experts Perception.

The sampling process is done for the selected variables chosen for modeling, splitting into training and testing dataset. The training data set are used to train neural network model and the testing data is used for the prediction of the data sets through the neural network trained model for the evaluation of SUITE PUE. Note that many of the inputs representing totals and averages are actually metavariables derived from individual sensor data.

Data preprocessing such as file I/O, data filtration and calculating metavariables, Variable Analysis was conducted using Excel, R. Both R and Matlab R2012a were used for model training, post processing and simulating results.

IV. Result And Discussion

The precise and robust PUE model offers many benefits for heterogeneous DC operators and owners. For example, in real time comparison of actual vs predicted heterogeneous DC performance for any given set of conditions can be used for automatic performance alerting, real-time plant efficiency assessing and troubleshooting.

A precise efficiency model also enables DC operators to evaluate PUE sensitivity to DC operational parameters. For example, an internal analysis of PUE versus Cold Aisle Temperature(°F) conducted at a heterogonous DC suggested a theoretical 0.0005 reduction in PUE by increasing the cooling tower LWT and chilled water injection pump set points by 3F. This simulated PUE reduction was subsequently verified with experimental test results after normalizing for server IT load and wet bulb temperature [5]. Such sensitivity analyses drive significant cost and carbon savings by locating and estimating the magnitude of opportunities for further PUE reductions.

Finally, a comprehensive DC efficiency model enables operators to simulate the DC operating configurations without making physical changes. Currently, it's very difficult for an operator to predict the effect of a plant configuration change on PUE prior to enacting the changes. This is due to the complexity of modern DCs, and the interactions between multiple control systems. A machine learning approach leverages the plethora of existing sensor data to develop a mathematical model that understands the relationships between operational parameters and the holistic energy efficiency. This type of simulation allows operators to virtualize the DC for the purpose of identifying optimal plant configurations while reducing the uncertainty surrounding plant changes.

A. *Prediction Results*

Figure 6 depicts a snapshot of predicted vs actual PUE values at one of heterogonous DCs over one month during the summer. The neural network detailed in this paper achieved a mean Square error of 0.004 and standard deviation of 0.001 on the test dataset. Note that the model error generally increases for PUE values greater than 1 .29 due to the shortage of training data corresponding to those values. The model

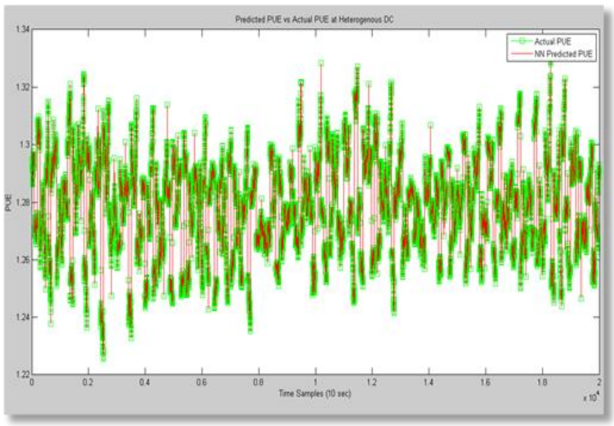

Figure 6. Predicted vs Actual PUE values at heterogonous DC

accuracy for those PUE ranges is expected to increase over time as additional data are collected on heterogeneous DC operations.

B. *Sensitivity Analysis*

The following graphs reveal the impact of individual operating parameters on the DC PUE. We isolate for the effects of specific variables by linearly varying one input at a time while holding all others constant. Such sensitivity analyses are used to evaluate the impact of set point changes and identify optimal set points. All test results have been verified empirically.

Figure 7a represents shows, as the Cooling coil leaving temperature (°F) increases, the PUE decreases. DC should be maintained with increasing the cooling coil leaving temperature with stabilizing other variables and making PUE more effective to reduce the cost. Similarly, Figure 7b suggests that the providing a system with cold aisle temperature (°F) over a period of time under different circumstance, the variation in the PUE is linearly increased as the cold aisle temperature decreases.

Figure 7c represents a linear variation as the Cooling coil valve position increases the PUE also increases, as it is directly proportional the usage of power is more as it becomes big. Figure 7d indicates that when Cold coil out water temperature decreases eventually the PUE increases, so the temperature for this scenario is optimized and they are inversely proportional to each other.

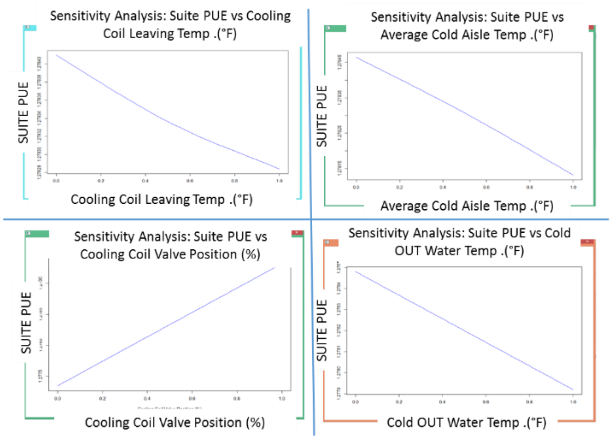

Figure 7a-7d: SUITE PUE vs Cooling Coil Leaving Temperature, Average Cold Aisle Temperature, Cooling Coil Valve Position and Cold Coil Out Water Temp

Figure 8a represents that as the cooling coil chilled liquid flow increases significantly the SUITE PUE decreases so there is an inversely proportional to each other.

Figure 8b represents a slightly sloppy curve for the SUITE PUE versus Heat reclaim coil leaving temperature (°F), says that PUE is in stabilized state when the temperature is in the optimal stage and also shows that they are inversely proportional as the temperature increases the PUE drops out.

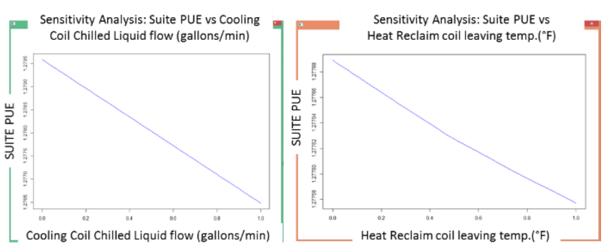

Figure 8a-8b: SUITE PUE vs Cooling Coil Chilled Liquid Flow and Heat Reclaim Coil Leaving Temperature

fan power for controlling the PUE without exceeding drastic change in the power consumption.

Figure 9a & 9b show that Fan Power and Fan Speed are directly proportional to SUITE PUE, where Figure 9a signifies a linear variation between the PUE and Fan Power but Figure 9b depicts that there is an optimization in fan speed through an upper sloppy curve which creates a positive impact in the

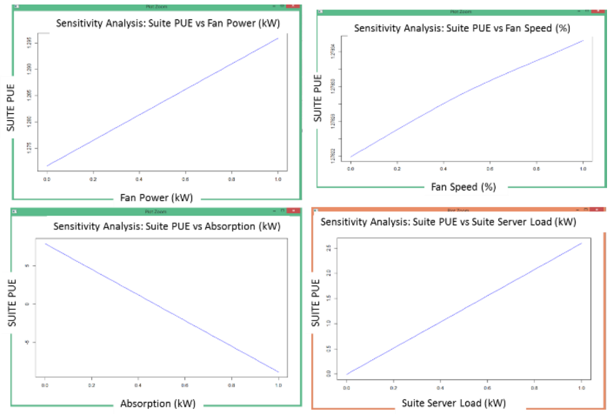

Figure 9a-9d: SUITE PUE vs Fan Power (KW), Fan Speed (KW), Absorption (KW) and Suite Server Load (KW)

Figure 9c signifies that the Absorption (KW) which is the chiller power varies inversely to PUE, as chiller power increases PUE drops. It concludes that it creates a great impact in PUE, which relatively stabilized through the fan power and server load for better synchronization.

Figure 9d specifies the variation of PUE with Suite Server Load (KW) is linear, which states that the PUE decreases exponentially as the server load decreases. Eventually as per the data samples trained most of the power in the heterogeneous DC station is consumed by server load 78%.

Figure 10 represents that the accuracy of the Neural Network model with test cases empirically verified in matlab simulation [16]. The variation of PUE from the actual calculation with Neural Network trained model gives optimized results.

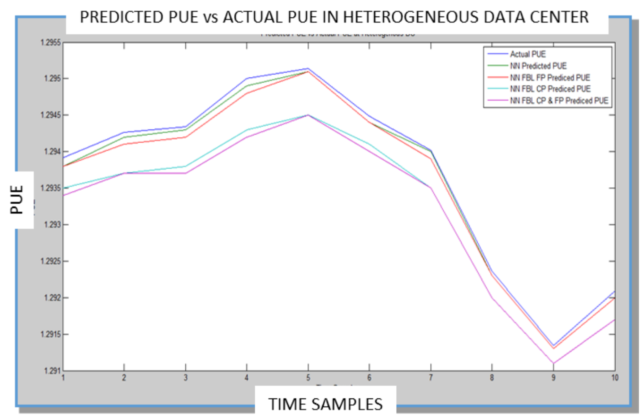

Figure 10: Neural Network Based Controller Output for DC

 ISBN: 978-1-61208-484-8

The model simulation done in four different scenarios for PUE optimization, such as Neural Network predicting PUE, Optimizing Fan Power (FP), which feedback to Neural network to Predict PUE, Optimizing Chiller Power (CP) which feedback to Neural Network to predict PUE and finally the best accuracy is obtained from optimizing both Fan Power (FP) and Chiller Power (CP).

Machine learning applications based on neural network based controller are limited by the quality and quantity of the data inputs. As one of important aspects to have a full range of DC operational conditions to precisely train the mathematical model. The model accuracy may decrease for conditions where there is less data. As with all empirical curves fitting, the same prediction results may be achieved for multiple model parameters θ. It is up to the analyst and DC operator to apply reasonable discretion when evaluating model predictions.

V. CONCLUSION

Accelerating growth in data center complexity and scale is making energy efficiency optimization increasingly important yet difficult to achieve. Though the model is simulated for heterogeneous data center environment where servers placed are of different kind, so the variation causes high end and low end rather than median. This made effective through machine learning and acquired best gain in PUE. Using the machine learning framework developed in this paper, we are able to predict DC PUE within 0.0004 +/- 0.0005. Using machine learning technique, you can further optimize power usage efficiency between 1% to 3%. This can translate is saving hundreds of thousand dollars in a datacenter. Actual testing on heterogeneous DCs indicates that machine learning is an effective method of using existing sensor data to model DC energy efficiency, and can yield significant cost savings. Model applications include DC simulation to evaluate new plant configurations, assessing energy efficiency, and identifying optimization opportunities.

ACKNOWLEDGEMENT

I would like to thank the leadership of a midsize data center in Indiana, USA to help offer their facility of this experiment.

REFERENCES

[1] Jonathan Koomey. Growth in Data centerelectricity use, 2005 to 2010. Analytics Press, Oakland, CA, 2011.

[2] Integrated Approach To Data Center Power Management Lakshmi Ganesh, Hakim Weatherspoon, Tudor Marian, Ken Birman Computer Science Department, Cornell University. 2012.

[3] IBM Interconnect 2015: A New Way. http://www.slideshare.net/ibm/ibm-interconnect-2015-asset-management-and-the-internet-of-things, May 2016

[4] http://www.data centerknowledge.com/archives/2011/11/15/pue-is-dead-the-case-for-performance-per-watt/ , May 2016

[5] Machine Learning Applications for Data Center Optimization Jim Gao, Google 2014.

[6] Impact of Hot and Cold Aisle Containment on Data Center Temperature and Efficiency Revision 2 by John Niemann, Kevin Brown, and Victor Avelar.

[7] Ramya Raghavendra, Parthasarathy Ranganathan, Vanish Talwar, Zhikui Wang, Xiaoyun Zhu." No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center".

[8] Automating Data center Operations Using Machine Learning by Peter Bod.

[9] Uptime Institute. Data CenterIndustry Survey. 2013.

[10] Frauke Gunther and Stefan Fritsch," neuralnet: Training of Neural Networks," The R Journal, Vol. 2/1, June 2010, pp 30-37.

[11] C. Marschner, "glm2: Fitting Generalized Linear Models with Convergence Problems," The R Journal, Vol. 3/2, December 2011, pp 12-15.

[12] Andy Liaw and Matthew Wiener, "Classification and Regression by randomForest," The R Journal, Vol. 2/3, December 2002, pp 18-22.

[13] Derrick , "Neural Network for self Learning control system," IEEE Control System magzine, Vol. 1, April 1990, pp 18-23.

[14] Swarup, K.S. and Subash, "Neural network approach to voltage and reactive power control in power systems" Intelligent Sensing and Information Processing, Proceedings of International Conference, 2005.

[15] Freeman J. A. and Skapura D. M.,. "Neural networks: Algorithm, Applications and Programming teaching techniques" AddisonWesley.1991.

[16] Chengming Lee and Rongshun Chen," Optimal Self-Tuning PID Controller Based on Low Power Consumption for a Server Fan Cooling System," Senors, Vol 15,pp 11685- 11700,2015.

[17] Deploying and using containerized/moular data center facilities, Christopher Kelley, Cisco Systems Jud Cooley, Oracle. https://www.thegreengrid.org/~/media/WhitePapers/WP42-DeployingAndUsingContainerizedModularData centerFacilities.pdf?lang=en, May 2016

[18] http://www.networkcomputing.com/data-centers/better-data-center-standardization-through-pod-architecture-design/55918907, May 2016

[19] http://dcblox.com/about-us/data-centers-2/ May 2016

[20] Research of PID Control Algorithm Based on Neural Network Liu Luoren, Luo Jinling ESEP 2011: 9-10 December 2011, Singapore