Artificial Neural Network Based Prediction of Control Strategies for Multiple Air-Cooling Units in a Raised-floor Data Center

Vibin Shalom Simon, Ashwin Siddarth, Dereje Agonafer
Mechanical and Aerospace Engineering
The University of Texas at Arlington,
P.O. Box 19023
Arlington, Texas, United States, 76019
Email: vibinshalom.simon@mavs.uta.edu

ABSTRACT

A data center cooling system consists of a hierarchy of systems with dedicated control algorithms dictating their operational states. There exists a wide range in spatial and temporal parameter space in an ensemble of non-linear dynamic systems, each executing a control task, while the global objective is to drive the overall system to an optimum operating condition i.e. minimum total operational power at desired rack inlet temperatures. Certainly, it is beneficial in optimizing workload migration at temporal scales but, solving the instability of the cooling systems operating at design points helps in understanding the whole system and make predictions to have better control strategies. Several techniques are available to realistically capture and make predictions. Data-driven modelling/Machine learning is one such method that is less expensive in terms of cost and time compared to other methods like validated CFD simulation/experimental setup.

The objective of this study is to develop a control framework based on predictions made using machine learning techniques such as Artificial Neural Network (ANN) to operate multiple Computer Room Air Conditioning Units (CRAC) or simply Air-Cooling Units (ACU) in a hot-aisle contained raised floor datacenter. This paper focuses on the methodology of gathering training datasets from numerous CFD simulations (Scenarios) to train the ANN model and make predictions with minimal error.

Each rack has a percentage of influence (zones) based on the placement of ACUs and their airflow behavior. These zones are mapped using steady state CFD simulation considering maximum CPU utilization and cooling provisioning. Using this map, ITE racks are targeted and given varying workload to force the corresponding ACU that is responsible for provisioning, to operate at set points. Number of such scenarios are simulated using the same CFD model with fixed bounds and constraints. Using large samples of data collected from CFD results, the ANN is trained to predict values that correspond to the activation of the desired ACU. Such efficient control network would minimize excessive cooling. The validated prediction points are used to model a control framework for the cooling system to quickly reach the operating point. These models can be used in real-time data centers provided; the training data is based on in-house sensor values.

KEY WORDS: Datacenter, Data-driven modeling, Air-cooling, control strategies, energy efficiency

Nomenclature

Q	Energy consumed by Air-cooling unit, J
M	mass flowrate of air, cfm
Cp	specific heat capacity of air, J/kg-K
N	number of iterations
T_{Return}	Temperature at the hot aisle, ^{0}C
T_{Supply}	Temperature at the cold aisle, ^{0}C
dT	change in temperature of air ($T_{Return} - T_{Supply}$), ^{0}C

Subscripts

Return	Return air at the ACU
Supply	Supply air provided by the ACU

Abbreviations

CFD	Computational Fluid Dynamics
ANN	Artificial Neural Network
ACU	Air-Cooling Unit
ITE	Information Technology Equipment
CRAC	Computer Room Air Conditioning
LMA	Levenberg-Marquardt Algorithm
CA	Cold Aisle
HA	Hot Aisle
LHS	Latin Hypercube Sampling
MSE	Mean Squared Error

INTRODUCTION

Data centers are facility buildings, housing Information Technology Equipment (ITE) and provide power and cooling. Technological advancement and price erosion enabled high growth rate of electronic packaging [1]. In 2014, data centers in the U.S. consumed an estimated 70 billion kWh, representing about 2% of total U.S. electricity consumption, and is estimated to increase to 75 billion kWh by 2020 [2]. A large data center at an industrial-scale operation uses as much electricity as a small town in the United States.

Managing cooling infrastructure is important to guarantee ITE reliability, working time, and operating scenarios for best performance. The scale of power consumption depends on workload, design and the longevity of data centers. Effective air distribution and provisioning of ITE will have a significant impact on energy consumption and equipment reliability[3]. The energy used by a typical rack of state-of-the art servers, consuming 20 KW of power at a cost of 10 cents per kWh is more than $17,000 per year in electricity. Data centers holding hundreds of such racks constitute an energy-intensive building. Efforts to improve energy efficiency in data centers can pay big dividends [4].

However, the cooling is generated and distributed by various systems, but airflow management is a key to optimum cooling of ITE and corresponding energy consumption. Optimizing the delivery of cool air and the removal of heat generated by the ITE can involve many design and operational practices. The general goal is to minimize or eliminate inadvertent mixing between supplied air to the ITE and hot air removed from it. Hot-aisle containment is one method to maintain the cold air supplied to the racks as generated by the cooling unit so that they are evenly distributed throughout the ITE without significant change in the temperature or humidity due to recirculation [4].

Generally, multiple Computer Room Air Control (CRAC) units respond altogether increasing cooling power to provision a localized hotspot which result s in excessive cooling for other ITEs especially in a co-located datacenter. The unnecessary cooling expenditure can be cut down by establishing a new strategy that could involve a particular CRAC unit or a combination of them to provision any localized hotspot. Knowing the fact that ITE workload fluctuation results in time-based temperature variation, the cooling unit also requires a certain time interval to respond to the scenario and make changes to address the situation.

Data center Facilities are of various types in terms of design, layout, ITE workload distribution and cooling strategies. Our data center design is chosen based on the literature survey from small-scale raised-floor data centers having indoor Computer Room Air Control (CRAC) Unit and hot aisle containment. The necessity of this design is to intentionally use the CRAC unit to provision localized hotspots due to workload distribution at any specific ITE so that the cooling power consumption can be optimized based on the need. The CFD model of the datacenter room does not involve Power distribution units, cables, pillars, exhaust vents and other supplementary equipment since they are considered insignificant in this study.

A robust CFD software, 6SigmaRoom provided by Future Facilities is used to model and simulate the temperature and flow characteristics based on a Blackbox model of ITE and several other units in a datacenter. The difficulties when using these softwares is that, it needs explicit domain knowledge and time expensive to produce the results for steady state/transient simulations. Since data centers are dynamic, CFD is not a suitable tool to produce real-time results to improve the power usage. Research has been conducted previously on such applications and it is found that Data-Driven Modeling (DDM) is viable option to analyze the data from the CFD model and validate with real-time raw data from the datacenter facility. Also, the methodology adopted to move forward in this research is based on the workflow demonstrated by Athavale et al. [5]. One such modeling technique is machine learning and the appropriate tool being used is Artificial Neural Network (ANN) to learn and mimic the behavior of airflow patterns and thermal characteristics in a datacenter. ANN has been used in various HVAC applications as well as in Datacenters for predicting parameters to control cooling unit based on the weather and psychrometric bins [6]. Predictions were made on different modes of cooling provided to datacenters depending upon the operational psychrometric bins and climatic conditions [7]. In our case the training dataset is generated by the CFD model having various configurations for cooling strategies for multiple Air-cooling units. The data driven model learns the non-linearity of physics-based systems and predicts parameters to modify the action space. The ANN is trained until it delivers the least error without overfitting the sample data, such that its prediction can be validated with in-house sensors deployed at specific locations in a datacenter facility.

Observing various configurations and types of data center we chose a model that is predominantly built in a small-scale raised-floor data center. The model is purposefully designed in such a way that the provisioning of ITE is visualized and quantified for various hotspot scenarios.

Raised Floor Hot Aisle Contained Data Center	
Total Room Capacity	300 KW
No. of racks per row	12
No. of Rows	3
Power per rack (KW)	8.4 (max.)
Cooling system	3 Air cooling Units (ACU) of 1140 KW max. sensible cooling

Table 1: Datacenter room specifications

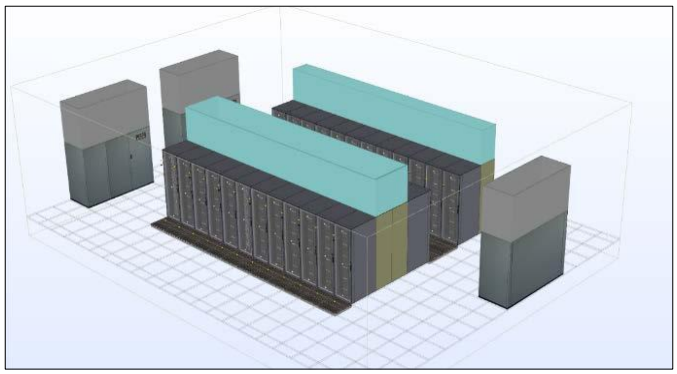
OBJECTIVES & STRUCTURE

The objectives of this study were to:

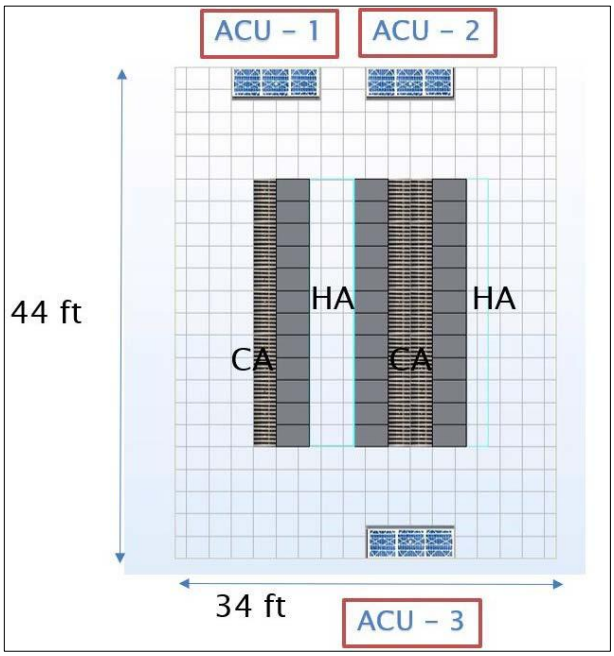
1. Understand the provisioning of ITE before an event of hotspot based on the zone of influence of the Air-cooling units over the servers.
2. Construct an ANN model using the CFD generated dataset for predicting temperature and airflow control parameters to operate the Air-cooling unit at desired operating points.

CFD MODEL & ITE SPECIFICATION

CFD analysis is carried out using the model showed in Fig.1. The model has underfloor supply and false-ceiling return configuration. The model has 2ft raised-floor design containing 36 racks, 12 racks per row provisioned by 3 ACUs. Ceiling is built at 14ft from the floor for the hot air to escape from the hot aisle containment to the return duct of the ACU. Solid obstructions are built from the hot-aisle containment to the vents in the ceiling to direct the hot air upwards. Similarly, the obstructions are built to direct the air from the ceiling to the return duct of the ACU. Floor grills of size (2 x 2) ft^2 with 50% open dampers are arranged in-line in front of the rack inlet to direct the cold-air upwards.



(a)



(b)

Fig.1 (a) Datacenter room model showing hot-aisle containment (b) Naming scheme for ACU

The underfloor plenum pressure is maintained using two sensors, one at the bottom of the tiles and the other above the rack. Usually these sensors are placed in the middle of the room.

Each rack is filled to its capacity with 1U servers of 200 W, a total of 42 servers per rack as shown in Fig. 2. Three Inlet and outlet temperature sensors are equally spaced along the air flow direction at the cold aisle and hot aisle respectively to capture the temperature stratification. Typical air leakages of 5% is set to all the racks. Initially, workload is distributed equally through all the servers, typical load of 40 W per ITE is given at idle conditions and 180 W per ITE is given at peak usage conditions. The above parameter is one such boundary conditions given during the simulation. The server used for modeling and analysis is HP SE1120 having an outflow pressure curve measured using experimental analysis using Air-Flow bench [8].

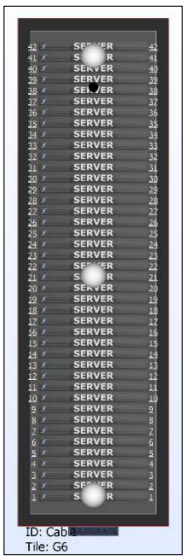


Fig. 2. Rack with 42, 1U servers and 3 equally spaced temperature sensors (white sphere)

The outflow pressure curve denotes the pressure difference across the ITE during its operation at various modes or workloads. It maintains an indirect relation with the cooling system performance. ITE power is time dependent and is set to fluctuate based on the workload distribution and migration.

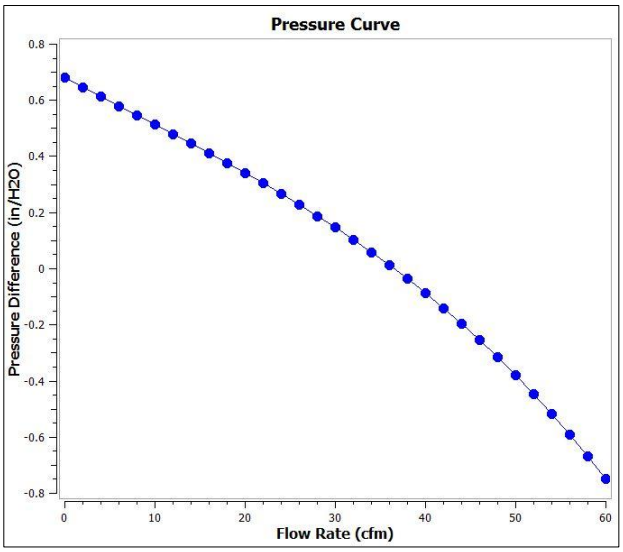


Fig. 3 Outflow Pressure Difference vs Air flow rate at the servers

AIR-COOLING UNIT CONTROL STRATEGIES

ACU uses chilled water-cooling system where the primary coolant is the water supplied from the chiller. The reference air i.e. Return air from the ITE is passed through the cooling coils to cool down to the required temperature. Supply temperature and flowrate variation is determined based on the thermal energy consumption equation embedded in the CFD software. Theoretically, the mass flow rate of air required to remove the heat generated by the ITEs can be calculated using the equation:

$$Q = m*Cp*dT$$

The Blackbox model of the servers are designed based on complex energy balance equations in CFD software that calculates parameters like temperature, pressure, velocity, humidity etc. The conventional construction of control network based on best practices for a small scale datacenter room are; Supply temperature of the air from the ACUs are controlled based on the temperature of air at the cold aisle, meaning, the work done/energy consumed by the heat exchanger depends on the server inlet air temperature. Supply air-flowrate is determined by the average pressure difference across the ITE in the datacenter, meaning, Variable Frequency drives are set to changing frequencies to operate the blower at different speeds to provide the desired air flowrate.

In this study, to understand the ACU's influence on provisioning the ITE's, we have setup the control network in such a manner that can be Average ITE outlet temperature sensor values based on % of influence are taken as T_{Return}. Similarly, supply temperature T_{Supply} setpoint is set to 22°C such that the ACUs respond when any of the inlet temperature sensors read a value of more than 22^0C. The blower speed for the ACU fans are controlled using VFDs and the values are updated for every iteration to capture different scenarios using same boundary conditions.

CRAC unit dimensions and specifications are modeled according to Liebert CW 114, ACU built by Vertiv cooling technologies.

Major ACU design parameters are listed below:

Total sensible cooling capacity: 114 KW

Max. coolant flow rate: 6 GPM

Supply air flowrate range: 0 – 17,300 CFM

Two CFD analysis were carried out, one to visualize the zone of influence of the CRAC units in provisioning the ITEs and the second to calculate the total power consumed by the CRAC units in various scenarios also to generate datasets for ANN training.

ZONE OF INFLUENCE ANALYSIS

ACUs supply air to the room through underfloor plenum and reaches the racks in a random fashion. To visualize and understand the influence of an ACU supply over the racks we simulate a steady state analysis with set boundary conditions given below.

Boundary conditions:

ACU Blower Speed: 90 % (15,570 CFM)

ACU Supply temperature: 22^0C

ITE Power/workload: 180 W

Once the spatial locations of the racks are found having maximum influence (75–100%) of every ACU, they are assigned as targeted ITEs. Power given to the targeted ITE is chosen such that it mimics the actual workload based on consumer usage. To address the fluctuating load on ITE, the ACU which has the highest influence on the respective servers start to respond. In a typical datacenter, all the ACUs respond together for minimal change in the workload but here, we forcibly allow the corresponding ACU to respond to the workload.

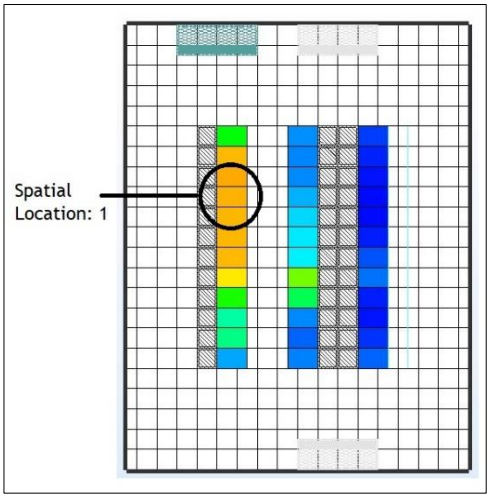


Fig. 4 Zone of Influence of ACU-1

From steady state analysis, ACU-1 has the highest influence on the racks at spatial location: 1. These racks are targeted and given fluctuating workload for a certain interval and ACU-1 is forced to respond while the other ACUs are constantly working on providing cooling to other racks. By running such simulations, we can find the energy consumed by ACU-1 to provision the targeted rack.

Similarly, the model is analyzed for similar scenarios according to the influence of ACUs. To capture the variability, several combinations of hotspot scenarios were created to generate training datasets for the Artificial neural network model.

Using temperature dependent control for the ACU for varying IT load is one such strategy practiced in data centers. Datasets are generated by collecting data from a set of simulations using PAC study in 6SigmaRoom.

PARAMETERIC STUDY

The objective function of this study is to capture the non-linearity of the physical phenomenon in the data center. A methodology for choosing the input parameter space out of 'N' number of measurable parameters is carried out using Latin hypercube sampling (LHS), a statistical technique where the domain of interest is filled with samples portraying the variability shown in original data.

The multi-dimensional parameter space should be space-filling and non-collapsing to ensure a good variability in simulations [9]. CFD simulations are deterministic in nature therefore, it is important that the input parameter space is determined using a LHS technique to avoid any bias and introduce required variability in the training data [5]. Latin hypercube sampling (LHS), a statistical method for generating a near-random sample of parameter values from a

multidimensional distribution, ensures that the ensemble of random numbers is representative of the real variability [10].

Using Latin Hypercube Sampling from a range of parameters, the input parameter space is generated to provide the maximum variability in the CFD simulation thus yielding datasets having good number of features. The CFD model is run for different combinations of inputs to generate training datasets. Python is used to generate the LHS from a list of input parameters to have a space filling design. From known parameter values and resolutions, the input parameter space is defined for different combinations for 27 CFD simulations.

Inputs for the CFD simulations given are: Time based ITE Load, Spatial Location of targeted ITE, power ratio for the ITE.

Time varying Outputs from the CFD simulations obtained are total cooling power consumption of all three ACUs in different scenarios and ACU blower speed.

Input Variables	Bounds	Constraints
ACU Blower speed	50 – 90 %	Interval of 20 (50%, 70%,90%)
Total ITE Workload / rack	4 – 8 KW	Interval of 2 (4,6,8)
Spatial Location	1 - 3	Interval of 1 (1,2,3)
Temperature rise of the IT load	5 – 10^{0}C	Interval of 1 (5,6,7,…10)

Table 2: Input parameter space

The two functions that govern the control parameters being captured and learnt mathematically by the ANN model are as follows:

1. ACU Blower Speed = f(ACU number , Spatial Location, Outlet Temperature, IT Load)
2. ACU Cooling power = f(ACU number , Spatial Location, Supply temperature, IT Load)

ARTIFICIAL NEURAL NETWORK

The sheer number of possible equipment combinations and their setpoint values makes it difficult to determine where the optimal efficiency lies [11]. Using standard formulas for predictive modeling often produces large errors because they fail to capture such complex interdependencies. Data driven models are the best methods that can completely represent a non-linear physics-based scenario in mathematical form so that we can train neural network models to learn and predict the desired parameters. Neural networks are a class of machine learning algorithms that mimic cognitive behavior via interactions between artificial neurons [12].

ANN TRAINING & VALIDATION

Using datasets from the CFD simulations, ANN model is trained, validated, and tested to predict the desired outputs that can allow us to frame a control strategy for the provisioning the ITE running under various workload scenarios. 70% of the dataset is used for training the neural network, the remaining 30% used for validation and testing. Data pre-processing such as sampling and data filtration is done using Python (PyCharm by JetBrains) in conjunction with the NumPy module and visualization using matplotlib module. MATLAB R2019a has predefined ANN structures to model train, validate, test and post-process.

In our study, the input parameters for the ANN are chosen in such a way that they can be measured directly from the data center facility. The probability of error becomes negligible. The list of input parameters for the ANN model are listed as follows:

1. ACU number
2. Spatial location of the targeted rack
3. Temperature at the cold aisle (^{0}C)
4. Rack IT Load (W)

The output parameters of the ANN model are considered in a way such that it can be used to frame a control strategy for every ACU.

The output parameters are as follows:

1. Blower Speed (cfm) for all 3 ACUs
2. Sensible cooling load on (Power consumed – KW) all 3 ACUs

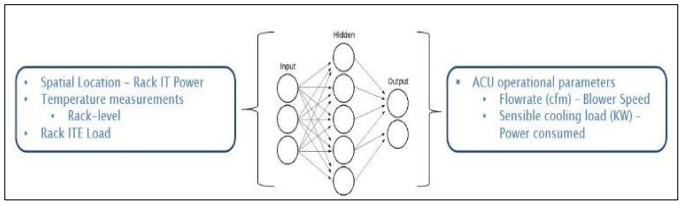


Fig. 5 ANN model

Number of neurons required to achieve minimal error in training the ANN is calculated with a set of values defined by the thumb rule [11].

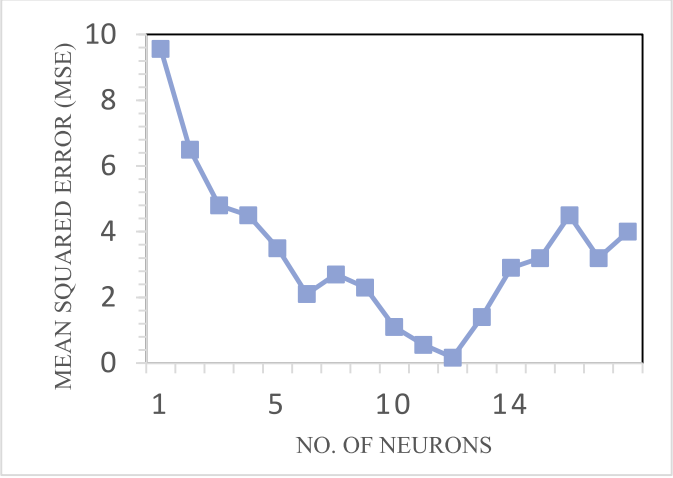


Fig. 6 ANN training error vs No. of neurons used in the hidden layer

In this case 11 neurons in the hidden layer were used for the ANN to predict results with minimal error. If we have neurons more or less than 11, we may have predicted results

with large error values also called as underfitting/overfitting of data.

The model selected was a three-layer model having one input, one output and one hidden layer. The dataset was trained using Levenberg-Marquardt Algorithm (LMA) to minimize the error as well as to overcome the flaws in using gradient descent method. Empirical relations are available to determine the suitable number of neurons for the hidden layer based on the number of parameters in the input and output layers [13-16]. The model is tested from a sample data that is not in the training dataset to evaluate the accuracy of the prediction. The Mean Squared Error (MSE) refers to the difference in original value and the predicted value, the lesser the value the more accurate is the prediction. Accuracy is improved by generating training dataset having higher resolution by increasing the number of scenarios using the CFD model.

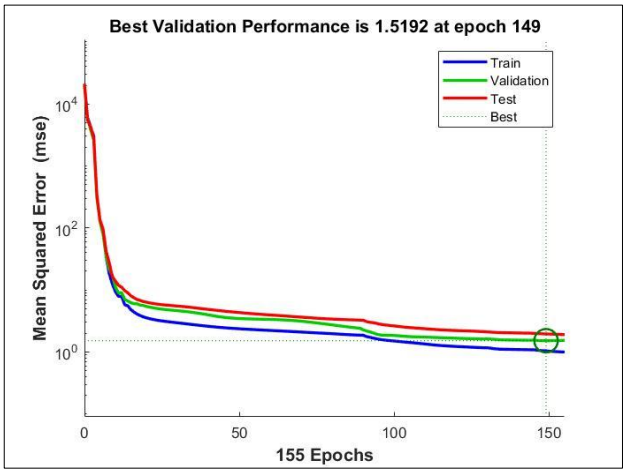


Fig. 7 ANN Training Performance

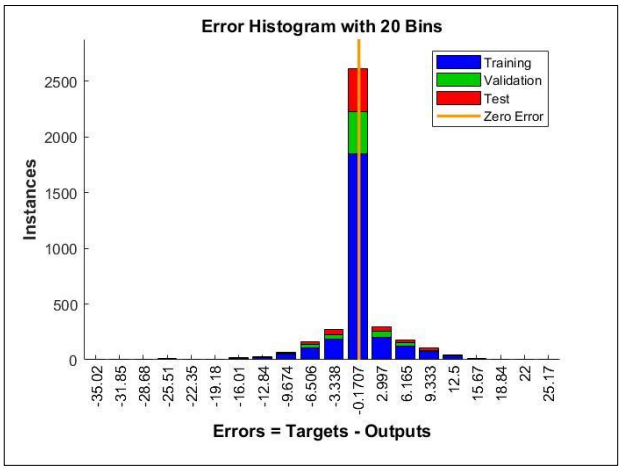


Fig. 8 ANN Training Error

From these results we can observe that ANN model can predict the parameters for ACU control with MSE in the order of 10^{-1} showing the model is a good fit to the data. All these parameters are fed into the ANN controller module that is integrated with the DC facility.

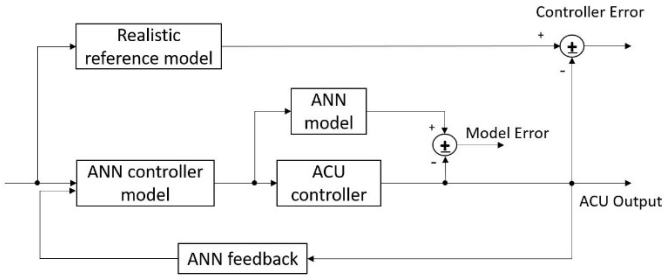


Fig. 9 ANN – DC controller network

The above network shows a one of the implementation techniques to have a control framework based on the ANN model thereby merging it with the datacenter facility. The feedback loop improves the ANN controller model for better prediction. Model error and controller error obtained from the comparison is used for further analysis and training of the system control design.

SUMMARY & CONCLUSIONS

This paper encapsulates an approach to address energy saving by using multiple ACUs in a datacenter using predictive modeling with training datasets collected using CFD simulations. Summarizing ANN test prediction results in an average error <3KW of energy consumed by the cooling units and 0.2% of the air flow using a part of the training data as test data.

The predictions are then used to design a framework that allows the operation of multiple ACUs to optimally provision the ITE load. ANN model resulted in a good agreement with the CFD model having error at the order of 10^{-1} when tested using a part of the training data can be used to frame a control strategy based on the hotspot in a typical raised floor data center with chilled water-cooling system. This ANN model can be implemented in a realistic data center provided; it is trained based on in-house sensor values. Cooling strategy can be essentially based on the ANN predicted results thereby reducing power consumed by the cooling units.

Following up this research could be, having high resolution CFD models and a greater number of hotspot scenarios to address excessive cooling provisioning, increasing the resolution of input parameter space to get more variability in the dataset also, cooling unit failure analysis can be included in predicting ACU operational parameters. Step ahead prediction of parameters is one such method where we will be able to have a better response from the cooling unit and suitable for real-time implementation as well.

ACKNOWLEDGMENTS

This research is supported by the National Science Foundation, Industry/University Cooperative Research Center and Energy Smart Electronic Systems (ES2).

REFERENCES

[1] Scaramella, J. "Next-Generation Power and Cooling for Blade Environments." *IDC, Framingham, MA, Technical Report* 215675 (2008).

[2] Koomey, Jonathan. "Growth in data center electricity use 2005 to 2010." *A report by Analytical Press, completed at the request of The New York Times* 9 (2011): 161.http://www.analyticspress.com/datacenters.html

[3] EPA. ENERGY STAR Rating for Data Centers: Frequently Asked Questions [Online]. Available: https://www.energystar.gov/ia/partners/prod_development/downloads/DataCenterFAQs.pdf?acac-cbed.

[4] Greenberg, Steve, et al. "Best practices for data centers: Lessons learned from benchmarking 22 data centers." *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings in Asilomar, CA. ACEEE, August* 3 (2006): 76-87.

[5] Athavale, Jayati, Yogendra Joshi, and Minami Yoda. "Artificial neural network-based prediction of temperature and flow profile in data centers." *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. IEEE, 2018.

[6] Adejokun, Feyisola, Siddarth, Ashwin, Guhe, Abhishek, and Agonafer, Dereje. "Weather Analysis Using Neural Networks for Modular Data Centers." *Proceedings of the ASME 2018 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems*. *ASME 2018 International Technical Conference and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems*. San Francisco, California, USA. August 27–30, 2018. V001T02A001. ASME. https://doi.org/10.1115/IPACK2018-8253

[7] Walekar, Abhishek, et al. "Neural Network Based Bin Analysis for Indirect/Direct Evaporative Cooling of Modular Data Centers." *ASME 2018 International Mechanical Engineering Congress and Exposition*. American Society of Mechanical Engineers Digital Collection, 2018.

[8] Betsegaw Kebede Gebrehiwot, "Maximizing use of air-side economization, direct and indirect evaporative cooling for energy efficient data centers", Diss. 2016

[9] Montgomery, Douglas C. *Design and analysis of experiments*. John wiley & sons, 2017.

[10] Stein, Michael. "Large sample properties of simulations using Latin hypercube sampling." *Technometrics* 29.2 (1987): 143-151.

[11] Gao, Jim. "Machine learning applications for data center optimization." (2014).

[12] Andrew Ng. "Neural Networks: Representation (Week 4)." Machine Learning. Retrieved from https://class.coursera.org/ml2012002.2012. Lecture.

[13] Mehrotra, Kishan, Chilukuri K. Mohan, and Sanjay Ranka. *Elements of artificial neural networks*. MIT press, 1997.

[14] Lippmann, Richard. "An introduction to computing with neural nets." *IEEE Assp magazine* 4.2 (1987): 4-22.

[15] Zhang, Lin, et al. "Multivariate nonlinear modelling of fluorescence data by neural network with hidden node pruning algorithm." *Analytica Chimica Acta* 344.1-2 (1997): 29-39.

[16] Daqi, Gao, and Wu Shouyi. "An optimization method for the topological structures of feed-forward multi-layer neural networks." *Pattern recognition* 31.9 (1998): 1337-1342.