E-ISSN: 2706-8927
P-ISSN: 2706-8919
www.allstudyjournal.com
IJAAS 2021; 3(1): 320-325
Received: 15-11-2020
Accepted: 23-12-2020

Manideep Yenugula
Kroger, Blue Ash, Ohio, United States

Corresponding Author:
Manideep Yenugula
Kroger, Blue Ash, Ohio, United States

Data center power management using neural network

Manideep Yenugula

DOI: https://doi.org/10.33545/27068919.2021.v3.i1d.1124

Abstract
Cloud based services have grown substantially due to the cost-effective migration of applications to the cloud. As a result, there are now a plethora of data centers that can provide these services on a massive scale, with a wide variety of the user experience and very little downtime. The need to control the power consumption and performance of the data center's constituent nodes without affecting service level agreements (SLAs) arises from the promise to provide differentiated services on a large scale. The efficiency of the power consumption in such data centers poses a significant challenge to cloud computing. Data center server energy consumption may be reduced by the use of various optimization methods, such as workload consolidation and machine location. Here, we provide a data-driven predictive neural network architecture that, at any given time in the future, can accurately predict the server's power consumption by taking into account all of its components in addition to the load of incoming requests.

Keywords: Energy savings, cloud computing, data centers, machine learning, neural networks, and optimization

1. Introductions
There is great promise for real-time near-optimal solutions with optimum predictive power management systems for hybrid electric cars. An accurate priori estimation of power consumption and the use of simplified driveline simulations to obtain an optimal control strategy live are key concerns for power management algorithm prediction. Finding suitable answers to these challenges is essential if real-time PMSs are to establish efficient power management strategies and encourage greater energy economy in electric powertrains [1]. Data centers may maximize their use of energy storage batteries to lower their Total Cost of Ownership, or TCO, as power storage methods for energy improve in capacity, cycle life, and reliability. With careful control of battery charges and discharges while preserving uninterruptible power systems capacity, data centers may reduce power consumption by filling the valley and lowering the peak [2]. To combat the climate with energy crises, a common strategy nowadays is to use a mix of traditional power plants and renewable power sources like solar panels. Energy management becomes much more complex when users' power demands fluctuate and solar systems exhibit unpredictable behavior [3]. Customers of cloud services are able to dynamically adjust their resource demands due to the flexibility of cloud resources. However, issues with resource utilization, load imbalance, with excessive power consumption might arise due to changes in resource needs and the pre-defined dimensions of VMs [4]. Hosting Internet-related services with cloud computing is made possible by the physical infrastructure provided by data center networks. Providers of Internet-related services and cloud computing must ensure the correct design of their data center networks in order to get competitive advantages in terms of service quality and cost efficiency [5]. In order to address the drawbacks of limited data, inflexibility of set mathematical models, inaccurate power predictions, and inept power management, a cloud-based system for new energy power prediction and control was developed. Together, the monitoring subplatform and the cloud subplatform constitute a twin platform that can simultaneously address the needs of cutting-edge big data creation and real-time control. At the same time, you may establish real-time connection to the power grid dispatching center. The model for predicting the electricity output of wind farms and solar systems is based on a neural network developed by BP [6]. The exponential growth in data volume has elevated networks of data centers to the status of essential infrastructure. Oversubscription, poor space utilization, excessive power consumption, and complicated cabling are some of the issues that DCNs face as a result of this growth [7].

2. Related Work

The report suggested using a feed forward neural network based on particle swarm optimization to estimate the aging of batteries in [8]. To fine-tune the FNN's biases and weights, that PSO is used. In order to ensure that the suggested technique is legitimate, it was compared against traditional FNN using battery data sets that were supplied by the NASA Prognostics Centre of Excellence. The simulation findings demonstrate that PSO-FNN performs much better in systems with a high degree of volatility.

Because of its accessibility, resources like as solar, wind, geothermal, as well as fuel cells are of utmost relevance from a [9] perspective. Tightening and synchronizing MGs to the power grid is very challenging because to their scattered and small-scale energy generation. Proper monitoring and management of electricity quality is important even after integrating MG with conventional grid systems. That research utilizes a Distributed Static Compensator trained by an Artificial Neural Network to monitor and maintain the Power Factor, active power, and reactive power of a grid-tied MG system. When put side by side with a conventional fuzzy logic controller, ANN proves to be the better choice. The suggested system's simulation data is used to conduct a short study under various operating conditions utilizing Matlab/Simulink architecture. Based on the outcomes, it is clear that ANN controllers outperform FLCs when it comes to enhancing system properties like stability, dynamic responsiveness, and efficiency.

The article demonstrates a fuel cell vehicle predictive PMS that is built on neural networks in [10]. In order to create the necessary PMS prediction models, the suggested technique employs two network types: time-delay as well as nonlinear autoregressive with external inputs. An ideal power split strategy taking into account low energy usage as well as on-board charge retention is investigated by the online control component across the expected horizon. A test driving cycle's global optimum solution and a rule-based technique are both taken into account for comparison assessment. The results showed that the suggested strategy might enhance energy efficiency by 20.71% without affecting the state-of-charge of energy storage devices.

Research will employ a Recurrent Neural Network approach with a Bayesian Regularization Algorithm to forecast PV power production one day in advance [11] due to the system's ability to resolve issues with prediction, classification, as well as energy management. The average absolute percentage error is used to determine the degree of inaccuracy in the study's simulation results. A comparison is made between the actual data and the accuracy of PV power forecasts using the RNN algorithm. In the future, the grid will make up for the amount of electricity that PV cannot provide. The optimal MAPE value of 2.2784% was achieved by predicting PV power using the RNN technique with four neuron hidden layers with a learning rate of 0.01. You may use the findings to predict how much PV power you will need for the next day by combining the RNN approach with past data.

The use of Deep Reinforcement Learning for controlling the charging and discharging of data center batteries is examined in the work [12]. The optimal charging and discharging times for batteries are determined by taking into account the current power price, the battery's state, and its cycle life. That allows for the optimization of savings on electricity bills. They meticulously planned the system's state, charging and discharge operations, reward function, and neural network architecture to get larger advantages. The results of the simulations show that the suggested algorithm can determine the optimal savings strategy in the energy pricing regimes of both the United States and Beijing. Priority experienced playback Deep Q-network may enhance energy storage savings by 47% and 55% with US and Beijing power costs, respectively, as compared to the baseline algorithm.

A novel adaptive learning network is suggested in the paper [13] by combining a neuro-fuzzy inference system that adapts network with a deep determinism policy gradients network. An innovative global K-fold fuzzy training technique is used to build the ANFIS network, which is then used to implement the offline dynamical programming solutions in real-time. The next step is to build the DDPG network so that the ANFIS network may regulate its input with the real-world reinforcement signal. By integrating the ANFIS and DDPG networks, a control utility that is dependent on the vehicle's energy consumption with battery state-of-charge may be optimized. Experimental examinations have validated the DDPG-ANFIS system's dependability and efficiency. The vehicle under investigation attained a CU that was 8% higher when fitted with a DDPG-ANFIS network as opposed to when the MATLAB ANFIS toolkit was employed. The DDPG-ANFIS network outperformed the ANFIS-only network by 138% and the DDPG-only network by 5% in five simulated real-world driving situations with the greatest mean CU value.

Using smart grids and artificial neural networks provide a novel approach to energy management in [14]. That will make sure that even if energy use and generation are operating at random, the customer will always have access to power. The MATLAB/Simulink tool is used to model and simulate the global system.

Data on total generated solar power in Turkey from 2009 to 2019 was used in the research [15]. They used this information to train an Artificial Neural Network (ANN) using Bidirectional long- and short-term memory techniques to predict the amount of solar electricity that would be installed in 2020. The cumulative power supply was predicted, and the results were assessed and comprehended.

3. Proposed Study

3.1 System Model

Neural networks are a kind of machine learning algorithm that find application in several domains. Some examples include sports scores, market pricing, weather forecasts, safety-related applications, voice and photo recognition, as well as other data-driven predictions. Neural networks are able to learn new information automatically by detecting patterns in the given training data. Neural networks outperform statistical models when it comes to interpreting information that is non-linearly varied and linearly dependant. When making a prediction, it's best to utilize both independent and dependent data. This can help prevent overfitting, which may happen when only using dependent data in a model. The accuracy improves in direct proportion to the size of the training data set.

3.2 Power Management Model

Efficient power administration in data centers is critical for minimizing energy consumption and cost, and this can only be achieved with an accurate host power model. Various

workloads and host conditions need evaluating the power models. We have tested several host setups and workload types (CPU, memory, as well as disk-intensive) to evaluate various current power models. Several performance counter characteristics that govern system power consumption have been established by an analysis of system performance and the nature of the hosts' power consumption.

As an example of publicly accessible experimental data, this article makes use of Google cluster workload. The workload in a data center is made up of jobs, and each job might have more than one task. Each task's start and finish times, as well as its CPU and memory utilization, are detailed in the Google cluster trace. This data allows us to determine the server cluster's (CPU as well as memory) resource utilization for each instance. Each time slot, which is usually 5 minutes long, has its CPU and memory resource utilization normalized to the maximum resource usage for the whole time frame. By combining precise server power modeling with CPU and memory use data, we are able to approximate the server cluster's power consumption. We use the predicted power consumption trace, the CPU and memory utilization traces, and other metrics as the experimental workload information.

Kindly be informed that the data center's power consumption forecast system has the potential to include the power use effectiveness (PUE) proportion, which mitigates the effects of various inefficiencies (such as cooling power). The following is the definition of PUE, which assesses the connection between the overall facility energy consumption and the energy used by IT equipment:

$$PUE = \frac{\text{Total Facility Power}}{IT\ \text{Equipment Power}} \quad (1)$$

By factoring in the PUE ratio, the prediction approach is able to go beyond estimating the power consumption of IT equipment and directly estimate the whole power consumption of the facility.

3.3 Neural Network with Q-Learning Model

When dealing with MDP issues, reinforcement learning is often the method of choice. When faced with challenges in an unfamiliar setting, reinforcement learning methods like Q-learning might be useful. The fundamentals of reinforcement learning as it pertains to data centers and power management are shown in Fig. Here, the agent's present state is determined by environmental input, and it changes depending on actions done and incoming workload. After that, we'll provide a quick rundown of Q-learning before introducing the deep neural network (DRL) based power allocation technique.

Using real-world operations data, we trained the structure of neural networks to predict the power consumption of distributed DCs.

First, we use a discrete-time Markov decision process (MDP) to describe our issue. In this approach, the whole time horizon is partitioned into time slots, which forms the basis of a reinforcement learning-based solution$= 0,1,\cdots,\infty$. The power demand is included in the system's state s^t at period tp_D^t, battery state b^t, and the temperature of the server's intake T_{in}^t as the power allocation defines the activity p_a^t. We may summarize our problem's MDP formulation as follows:

- System state: $s^t = (p_D^t, b^t, T^t) \in \mathcal{S}$
- Action: $p_a^t \in \mathcal{A}(s^t)$
- State transition probability: $P(s^t, p_a^t, s^{t+1})$
- Reward function: $r^t = R(s^t, p_a^t, s^{t+1})$
- Discount factor: $\gamma \in (0,1)$

The present state s^t determines the action area A in this case. The pair (s^t, p_a^t, s^{t+1}) results in a change in the system's status s^t to s^{t+1} when action p_a^t is occupied. Potential for a state to change $P(s^t, p_a^t, s^{t+1})$ gives the likelihood that the system's state will change from s^t to s^{t+1} assumed action p_a^t is in use at s^t. The reward task $R(s^t, p_a^t, s^{t+1})$ specifies the prompt benefit in the context of the state-action tuple (s^t, p_a^t, s^{t+1}).

Action Space. We examine a continuous action area for power allocation using two separate scenarios in our MDP formulation. We begin by allocating enough power to meet the total demand in cases when the data center's power capacity is insufficient, i.e., $p_a^t = p_D^t$. Secondly, we may add battery power up to its full capacity when demand is higher than capacity i.e., $p_a^t \le C_0 + \min(p_D^t - C_0, b^t)$. The present system states^t may be used to define the acceptable action space $\mathcal{A}$.

$$\mathcal{A} = \begin{cases} p_a^t = p_D^t, & \text{when } p_D^t \le C_0 \\ p_a^t \le C_0 + \min(p_D^t - C_0, b^t), & \text{when } p_D^t > C_0 \end{cases} \quad (2)$$

Q-learning. This model-free MDP problem-solving algorithm uses off-policy reinforcement learning. Put another way, Q-learning can successfully figure out the best course of action even when given no background information about its surroundings. For each potential state-action pair, the learnt strategy is stored as a discrete value of Q able. If we choose the action with the greatest value of Q in Eqn. (3), we can then extract the Q policy as π^Q

$$\pi^Q(s^t) = \underset{p_a^t \in \mathcal{A}(s^t)}{\operatorname{argmax}} Q(s^t, p_a^t) \quad (3)$$

Estimating Jf values for Q from the reaction of the environment is the key process in Q-learning. Through the use of a fixed-point loops of the Bellman equations (Eqn. (4)), the Q values for MDP in a specific environment may often be trained offline. It is possible to demonstrate the conventional Q-learning approach using a rate of learning α next.

$$Q(s^t, p_a^t) = Q(s^t, p_a^t) + \alpha\left[r^t + \gamma Q\left(s_{t+1}, \pi^Q(s_{t+1})\right)\right] \quad (4)$$

On state transitions with Markovian assumptions. The server inlet temperature and battery level b^t are relevant to our issue. T_{in}^t each changes as a result of the activity done to distribute electricity. Since the Markovian process is appropriately followed, the state of the battery b^t is only altered when the power transmission action p_a^t exceeds

the capacity C_0 as well as necessitates a battery supplementary to fulfill the demand. But our problem formulation needs a large enough temporal granularity Δt for temperature changes to happen and satisfy the MDP criterion. The reason for this is the time-consuming nature of changing the temperature. Nevertheless, this constraint on Δt may be circumvented by the use of an improved multi-level MDP, in which the subsequent state is influenced by both the present and previously visited states with actions; nevertheless, this expansion of the state-action field comes at a cost.

Contrarily, the MDP assumption is violated since modifications to the power consumption p_D^t primarily rely on user behavior as well as may not be related to the present condition and activity. The MDP formulation is made easier by taking the power requirement of the following time slot into account p_D^{t+1} is known at the time t by estimating the workload. Based on the present state and the action, our solution may identify the future state in this way. In order to automate the process of power demand estimate, we include an LSTM network into our design.

Reward function. Here is how we come up with our incentive function to achieve OPA's optimization goals:

$$r^t = R(s^t, p_a^t, s^{t+1}) = -L(p_D^t, p_a^t) - \beta_1 (T_{in}^{t+1} - T_{th})^+ \quad (5)$$

Where $(T_{in}^{t+1} - T_{th})^+ = \max(T_{in}^{t+1} - T_{th}, 0)$ constitutes a temperature infraction, $b^t - b^{t+1}$ the amount of power used by the battery during time slot t, along with the weight parameters are β1 and β2.

Similar to OPA's optimization goal, (5) rewards a reduction in latency favorably. Also, since both the energy from batteries and the temperature of the air can be remembered, we use penalties for overusing the batteries and temperature violations to include their effects into future choices. The optimization aim may be altered by adjusting the values of β_1 and β_2, which also function as units onversion coefficient ^1. As an example, a more conservative battery state will be achieved by raising β_2, whilst a more restricted temperature iolation state would be produced by arger values of β_1. Keep in mind that our action space satisfies the third power allocation restriction $\mathcal{A}(s^t)$. Finding the best action policy A^* to maximize the long-term payoff is the goal of the MDP issue $\sum_t \gamma \cdot r^t$ through a discount actor $\gamma \in (0,1)$. Here, we introduce the element of discounting γ in order to create an issue that can be handled easily.

A. Model Implementation As shown in figure, there is a general four-layer neural network. This network takes as input a matrix x with dimensions (m x n). The number of factors, like server load, temperature, power consumed by inter-process communication, etc., and the number of training samples, m and n, respectively, are used here. The information is received by the input layer of the sensory nodes. The network's fundamental units are neurons. Neuron activation function is a non-linear function that is realized by each neuron by adding the results of weights coefficients with input data. The expected power, denoted as hθ(x), is the result of this process occurring at each layer. The complexity of the system determines the size and quantity of neurons. The neural network must be trained following four distinct phases.

Using a non-linear activation function called a Smooth Rectangle Unit (ReLU), we trained an ANN, or artificial neural network, that uses feed-forward back propagation. The ANN's input and weight vectors were regularized via this process. Equation (6) defines the Smooth ReLU:

$$f(x) = \begin{cases} x & \text{if input layer.} \\ \frac{e^x}{1+e^x} & \text{otherwise.} \end{cases} \quad (6)$$

When assessing the model's performance across unknown patterns, we additionally took into account the Root mean square error (RMSE) and the Mean Absolute Percentage Errors (MAPE), as shown in equations (7) and (8):

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(f(x_i) - y_i)^2} \quad (7)$$

$$MAPE = \frac{100}{M}\sum_{i=1}^{M}\frac{|f(x_i) - y_i)|}{y_i} \quad (8)$$

Where $f(x_i)$ and y_i denote the expected and observed values, correspondingly, and M stands for the quantity of training data. Here is how we calculated the accuracy percentage:

$$\text{Accuracy } (\%) = 1 - \text{MAPE } (\%) \quad (9)$$

4. Results and Discussion

A cloud with four geographically dispersed data centers (DCs)-DC1, DC2, DC3, and DC4-in Manitoba, Quebec, Ontario, and Minnesota, respectively-was proposed. In order to maintain the thorough surveillance, we additionally supplied our virtual DCs with several types of RE. Table 1 displays the details of our DCs, including the number of servers, server type and specifications, and more.

Table 1: Different Servers in DCN for Configuration

Location	Server type	Server number	Server spec.	
			Active(kW)	Idle (kW)
Manitoba	Intel E5506	1500	0.419	0.146
Quebec	Intel X5570	1000	0.352	0.153
Minnesota	AMD EPYC 7601	2000	0.483	0.138
Ontario	Intel E5-2699	1500	0.529	0.102

Table 2: ANN- along with LSTM-based predictions using actual data, together with average normalizing values of CPU, memory, and power consumption.

	CPU	Memory	Power
ANN	0.307	0.390	0.357
LSTM	0.286	0.437	0.354
Actual	0.302	0.435	0.362
ANN MAE	6.54%	7.35%	6.10%
ANN-QL	6.04%	3.53%	4.42%

As a last step in training for optimal accuracy, we use back propagation to alter the weights of the coupled neurons. In order to address the back propagating mistakes, we used an approach for adaptive rate of learning optimization called the Adam optimization. It is evident that our model has minimum computational complexity due to (i) the limited

quantity of neurons in the hidden layer and (ii) low-frequency of prediction execution.

In our ANN with QL technique, we began utilizing neural network weights distributed uniformly among [1, −1] to limit the error and avoid the formation of an unstable equilibrium. Due to the error's propagation backwards across the disguised layers, it is vital to apply the same settings. Lastly, our training information set contains 14 days' worth of operational data, or 12096 input samples, with a precision of 15 minutes each. Out of the complete data, we trained with 80% and tested with the remaining 20%.

Table 3: Performance Metrics using ANN-QL

Metric	Training/total data			
	90%	**80%**	**70%**	**60%**
RMSE	62.2	65.2	80.1	95.7
MAPE	22.8	25.3	28.2	29.5
Accuracy	95.2	95.56	96.2	97.5

Here, we check how well the prediction model worked. As previously stated, we used two measures for evaluation: RMSE and MAPE. To ensure accurate predictions, we used test data in conjunction with the training information to train the prediction model. The relationship between the amount of training information and the total data is shown in Table 7, along with the RMSE and MAPE.

The findings reveal that the prediction accuracy rises with higher training data sizes, while the result is reliable regardless of the amount of the training data. With an 80% training data quantity, Figure 8 shows how well our prediction model performed in both the training and evaluation procedures. On an Intel(R) Pentium G2120 CPU with 4 GB of RAM, we execute the 800 epochs using 498s to build the model.

5. Conclusion

This study's findings highlight two key points: first, the influence of all factors other than load on power prediction accuracy, and second, the capacity of neural networks to accurately extract relationships between heterogeneous components for power prediction. Additionally, the aggregater's continual monitoring, which is based on the established rules to take action on the anticipated power, aids in power containment at servers and, by extension, impacts total energy consumption. There is a high degree of agreement between the projected and actual power consumption here. In order to determine what steps to take next, the aggregator takes into account the expected power at each node. The paper's occurrences and the steps to be executed in response to them are specified by hand for the purpose of predicting power values. The end objective of our project is to use reinforcement learning mechanisms, such as Q-learning, to accurately describe events and their accompanying actions dynamically based on past inputs.

6. Future Work

We are also going to investigate a wider range of historical data. Moreover, leveraging the result of this study for proposing new approaches of power management in geo-distributed DC as well as creating new opportunities for participating in ancillary energy markets are among the goals that we left for future work.

7. References

1. Liang Y, Lu M, Shen ZM, Tang R. Data Center Network Design for Internet-Related Services and Cloud Computing. Production and Operations Management. 2021;30:2077-2101.
2. Zhang S, Li G, Li T, Han X, Ren Y, Kang J. Research on Power Control and Prediction Strategies of New Energy Based on Cloud Platform and BP Neural Network. 2021 International Conference on Power System Technology (POWERCON); c2021. p. 953-957.
3. Butun B, Onur E. Joint Virtual Machine Embedding and Wireless Data Center Topology Management. 2021 17th International Conference on Network and Service Management (CNSM); c2021. p. 63-69.
4. Kumar R, Khatri SK, Diván MJ. Power Usage Efficiency (PUE) Optimization with Counterpointing Machine Learning Techniques for Data Center Temperatures. International Journal of Mathematical, Engineering and Management Sciences; c2021.
5. Shojaeighadikolaei A, Ghasemi A, Bardas AG, Ahmadi R, Hashemi M. Weather-Aware Data-Driven Microgrid Energy Management Using Deep Reinforcement Learning. 2021 North American Power Symposium (NAPS); c2021. p. 1-6.
6. Saxena D, Singh AK, Buyya R. OP-MLB: An Online VM Prediction-Based Multi-Objective Load Balancing Framework for Resource Management at Cloud Data Center. IEEE Transactions on Cloud Computing. 2021;10:2804-2816.
7. Junhuathon N, Sakunphaisal G, Chayakulkheeree K. Li-ion Battery Aging Estimation Using Particle Swarm Optimization Based Feedforward Neural Network. 2020 International Conference on Power, Energy and Innovations (ICPEI); c2020. p. 73-76.
8. Acharya DP, Nayak N, Choudhury S. Active Power Management in an AC Microgrid Using Artificial Neural Network and DSTATCOM. 2021 International Conference in Advances in Power, Signal, and Information Technology (APSIT); c2021. p. 1-6.
9. Ali AM, Yacoub MI. Optimal predictive power management strategy for fuel cell electric vehicles using neural networks in real-time. 2020 IEEE Vehicle Power and Propulsion Conference (VPPC); c2020. p. 1-6.
10. Kusuma V, Privadi A, Setya Budi AL, Budiharto Putri VL. Photovoltaic Power Forecasting Using Recurrent Neural Network Based On Bayesian Regularization Algorithm. 2021 IEEE International Conference in Power Engineering Application (ICPEA); c2021. p. 109-114.
11. Yan L, Liu W, Jiang W, Li Y, Li R, Hu S. Deep Reinforcement Learning based Optimization of Battery Charging and Discharging Management for Data Center. 2021 International Joint Conference on Neural Networks (IJCNN); c2021. p. 1-9.
12. Zhou Q, Zhao D, Shuai B, Li Y, Williams H, *et al*. Knowledge Implementation and Transfer With an Adaptive Learning Network for Real-Time Power Management of the Plug-in Hybrid Vehicle. IEEE Transactions on Neural Networks and Learning Systems. 2021;32:5298-5308.
13. Jarmouni E, Mouhsen A, Lamhammedi M, Ouldzira H. Energy management system and supervision in a micro-

grid using artificial neural network technique. International Journal of Power Electronics and Drive Systems (IJPEDS); c2021.
14. Özdemir MH, Ince M, Aylak BL, Oral O, Taş MA. Installed solar power prediction for turkey using artificial neural network and bidirectional long short-term memory. Business & Management Studies: An International Journal; c2020.