P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

APPLICATION OF MACHINE LEARNING TECHNIQUES IN DATA CENTRE ENERGY MANAGEMENT

R. Ranjana[1], N. Susila[2], T. Subha[3]
[1,3]Department of Information Technology, Sri Sairam Engineering College
Chennai, Tamil Nadu
[2]Department of Information Technology ,Sri Krishna College of Engineering and Technology, Coimbatore E-mail: ranjana.it@sairam.edu.in[1], susila@skcet.ac.in[2], subha.it@sairam.edu.in[3]

ABSTRACT: Cloud computing is one of the leading computing paradigms that offers services like Infrastructure as a Service called IaaS, Platform as a Service called PaaS, Software as a Service called SaaS to users on a pay per use model. The massive data centers that help cloud offer all the above stated services are virtualized. Virtualization enables easy management of resources. However, the massive physical servers in the data centers tend to consume enormous energy, leading to high environmental impact. So, energy conservation with optimum usage and management is one of the prominent areas of research in cloud. The major techniques to manage energy is to identify unused physical resources and put them to low power state or sleep state. But, the usage of resources depends heavily on the user requirements in an elastic environment like cloud. Hence machine learning techniques can be used to predict the usage patterns thereby identifying the physical resources required to fulfill the user demand. This paper aims to survey the avenues wherein machine learning can be applied to help energy management in a cloud data center.

KEYWORDS: data centres, cloud servers, machine learning, virtualization.

I. INTRODUCTION

With the rapid increase in the number of users over the internet,the need for more data and higher level of services via internet is also surging. This results in a need for improved capacity for data centers to handle large number of requests simultaneously. Extending the hardware resources is definitely the only way, but studies show that the basic networking equipment are not used to their fullest potential,which leads to substantial power wastages. Therefore, improving the energy consumption levels is the need for any data center.

An ideal energy conservation approach must take into account all the resources (server components, network devices, storage devices and cooling devices). The goal here is to achieve lower Power Usage Efficiency (PUE) index and at the same time provide quality services to all the users. The Cloud service providers offer services to many applications and therefore need to maintain service level agreements (SLAs), work under low access latencies and provide a secured service to the customers.

Data Centre energy optimization is definitely a critical mission to consider due to various dynamic factors like traffic patterns, workload distribution, resource mapping, etc. However,

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

the service providers are in a fix to consider these problems due to rapid increase in power consumption and the competitive price market for cloud services.

II. ENERGY MANAGEMENT IN DATA CENTER NETWORKS

All the network devices, irrespective of their level of utilization, are energy-hungry. Data Centers aim to connect a node with every other node on the same network. This has been implemented in hierarchical form in all Data Centers, with multiple sub-hierarchies connected via switches. These network devices, in spite of being less in numbers when compared to the servers, consume a considerable amount of power, since router peak power can be up to 90 times greater than server.

Power consumption in Data Centers can be managed using an Adaptive Link Rate (ALR) technique. This technique consists of a mechanism that defines link rate synchronization and a policy to control the link data rate. Its aim is to achieve lower data rates, lower power/idle (LPI) state transitions. ALR auto-negotiation tool, or the MAC frame handshake is used to synchronize the data link rates. This method has been effective in reducing power consumption to about 10% from the initial 90%.

Yet another issue is oversubscription. A need to reach hundreds of servers for a single search query is mandatory in most cases. In fact, the intra-data center communications accounts for a whopping 70% and these
require minimal latency. To solve this issue, server-centric and hybrid network architectures have been adopted; and 100-Gb Ethernet solutions have been proposed.

2.1. Need for Energy Management

All the modern servers operate at 10%-50% of maximum possible utilization. Despite the fact that average utilization remains very low, frequent bursts of activity can occur out of the blue. The need to meet the requirements of Service level Agreements (SLAs) forces the operators to allocate high amount of resources, which leads to poor energy efficiency. The design of green solutions for modern data centers has become a topic of paramount importance.

Power management should be taken care of in order to manage the expenses. Studies show that in near future, there may be a situation where the cost of power would surpass the cost of the actual setup. A high ratio of cooling power to computing power restricts the compaction and consolidation possible in data centers, which also increases the operation costs. For example, the high-power density poses significant challenges in routing the large amounts of power needed per rack. Currently the power delivery in typical data centers is near 60 Amps per rack and it is expected to reach the limit of power delivery, which will severely affect the operation of servers.

Maintaining the resources at an optimal temperature is also a key task to consider. It has been observed that a 15-degree Celsius rise increases the failure rates in hard-disks by a factor of two. Thus, for reliable and prolonged usage of the server, temperature maintenance plays a vital part. Further, the carbon emissions from such large data centers should also be reduced.

Hence, in this paper, we have conducted a survey on techniques for managing power consumption in data centers.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

2.2 Avenues foe Energy Consumption Reduction

2.2.1 Network device power consumption reduction

A Network Designer is the one who must consider information flows while designing the network architecture. The main idea here is to reduce the overload on the whole network. To achieve this, the inter-group communications is minimized and the intra-group communications are maximized. One very popular method to achieve this network scalability is the method of clustering [1]. Clustering algorithms have been used to identify idle groups for hibernating or switching them off.

In spite of many clustering algorithms in existence, not many of them work to reduce the energy consumption in network architecture [2]. However, many have been proposed that focus only on wireless sensor networks. The main objective of these algorithms is to reduce the number of clusters. These algorithms, which are heuristic in nature, are generally considered when the application is dependent on the length of the routing paths and varies based on data latencies. The idea of spectral algorithms, which work faster for large sparse graphs, was proved to be effective as well as efficient at finding the best possible clusters and the approximate number of clusters from real world data sets. Several literatures are available that discusses the different approaches and the advantages and disadvantages of different spectral clustering algorithms.

2.2.2 Physical Machine Reduction

One of the most sought-after ways to reduce the physical machine at data centers is implementing the concept of Virtualization. In spite of having many advantages, it falls short in the following ways:

Upfront costs:

The investment in the virtualization software, and possibly additional hardware might be required to make the virtualization possible. This depends on the existing network. Many businesses have sufficient capacity to accommodate the virtualization without requiring a lot of cash. This obstacle can also be more readily navigated by working with a Managed IT Services provider, who can offset this cost with monthly leasing or purchase plans.

Software licensing considerations:

This is becoming less of a problem as more software vendors adapt to the increased adoption of virtualization, but it is important to check with the vendors to clearly understand how they view software use in a virtualized environment.

Possible learning curve:

Implementing and managing a virtualized environment will require IT staff with expertise in virtualization. On the user side a typical virtual environment will operate similarly to the non-virtual environment. There are some applications that do not adapt well to the virtualized environment – this is something that your IT staff will need to be aware of and address prior to converting.

Therefore, it is more advantageous to introduce the concept of machine learning to reduce the physical machines in data centers. Google is one of the pioneers in implementing Data Centre energy efficiency models based on neural networks [4]. Neural networks are a class of machine learning algorithms that mimic cognitive behavior via interactions between artificial neurons [5]. Neural Networks does not need the user to predefine interactions in the model, which identifies the relationship between different classes of data. This is an advantage in case of modelling intricate systems. The neural network analyses for interactions between features to generate best-fit model. Here, the model accuracy increases with the due course of time.

The neural network utilizes 5 hidden layers, 50 nodes per hidden layer and 0.001 as the regularization parameter. The training dataset contains 19 normalized input variables and one normalized output variable (the DC PUE), each spanning 184,435-time samples at 5-minute resolution (approximately 2 years of operational data). 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. 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:

znorm= (z − MEAN(z))/(MAX(z) −
MIN(z)) The neural network features are listed as follows:

1. Total server IT load [kW]
2. Total Campus Core Network Room (CCNR) IT load [kW]
3. Total number of process water pumps (PWP) running
4. Mean PWP variable frequency drive (VFD) speed [%]
5. Total number of condenser water pumps (CWP) running
6. Mean CWP variable frequency drive (VFD) speed [%]
7. Total number of cooling towers running
8. Mean cooling tower leaving water temperature (LWT) set point [F]
9. Total number of coolers running
10. Total number of dry coolers running
11. Total number of chilled water injection pumps running
12. Mean chilled water injection pump setpoint temperature [F]
13. Mean heat exchanger approach temperature [F]
14. Outside air wet bulb (WB) temperature [F]
15. Outside air-dry bulb (DB) temperature [F]
16. Outside air enthalpy [kJ/kg]
17. Outside air relative humidity (RH) [%]
18. Outdoor wind speed [mph]
19. Outdoor wind direction [deg]

Note that many of the inputs representing totals and averages. Data pre-processing such as file I/O, data filtration and calculating metavariables was conducted using Python2.7 in conjunction with the Scipy 0.12.0 and Numpy 1.7.0 modules. Mat lab R2010a was used for

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

model training and post processing. Fig.1 shows the predicted vs. actual PUE values:

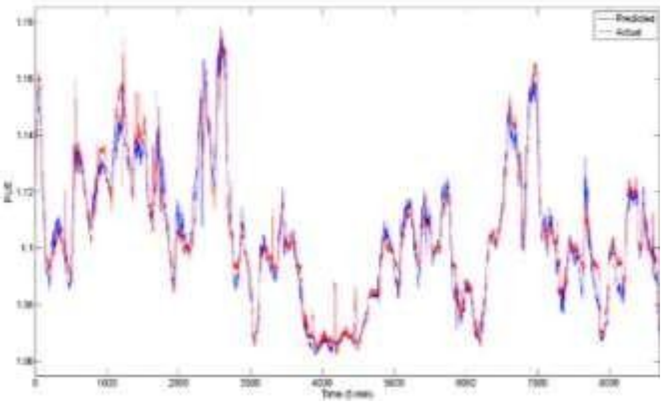

Fig.1: Predicted vs. Actual PUE values

The neural network detailed here achieved a mean absolute error of 0.004 and standard deviation of 0.005 on the test dataset. Note that the model error generally increases for PUE values greater than 1.14 due to the scarcity of training data corresponding to those values. The model accuracy for those PUE ranges has been expected to increase over time as Google collects additional data on its DC operations.

2.3 Cooling cost reduction

Of the total power consumption at a data center, studies show that almost 30% to 55% accounts for the cooling and ventilation systems. With rapid growth of large data centers worldwide, data centers become energy intensive processes accounting for over 1% of the world’s electricity usage. Large data centers with capacity up to 120 MW have been built in recent years. Energy efficiency becomes an issue of utmost importance for these data centers. Investigation showed that energy consumption by cooling data center IT equipment is between 30% and 55% of the total energy consumption.[6] Cooling and ventilation system consumes an average 40% of the total energy consumption in a data center.

In the conventional data center with hot aisle/cold aisle, cold air generated by the cooling system is supplied through a plenum under the floor and perforated air flow panels. The cold air flows up horizontally entering the tiny spaces between the servers from one side of the servers and leaving from another side [7]. Higher flow pressure drop, cold and warm air mixture on the upper side of racks are the main disadvantages. The concept of distributed air flow control is to divide a data center in pre-defined zones and different amounts of air flow are supplied to these zones across the data center based on local cooling loads. It requires special ventilation system design under the raised floor to distribute cooling air through ventilation ducks anddampers. However, there is always a risk of hot air and cold air mixing, which can cause considerable energy wastage.

In [8], the author presents the end-to-end cooling control algorithm (CCA), adapted from the DDPG, which combines the critical RL techniques and methods such as deep Qnetwork (DQN), deterministic policy gradient (DPG) and actor-critic algorithm. The proposed algorithm could predict the temperature control setting, which had shown a reduction of

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

about 15% of the total cooling cost.

2.2 Network topology-based improvement

One of the arenas which can further improve the scope of energy conservation in data centers is the topology of the network. It has a remarkable impact on the modularity and agility of the data center as a whole. Currently, data center networks use top of rack (ToR) switches that are interconnected through end of rack (EoR) switches, which are in turn connected via core switches. A significant over utilization of bandwidth in the network cores is the result of such an arrangement.

General topologies are generally classified into two: fixed architectures and flexible architectures. Fixed architectures can be further grouped into Fat Trees, Clos Network and recursive topologies such as DCell, BCube. Flexible architectures include Helios, cThrough and OSA.

Several analyses have been conducted in identifying the pros and cons of a network architecture. For instance, Rastin Pries, Michael Jarschel, Daniel Schlosser, Michael Klopf, and Phuoc Tran-Gia have done the research based on certain parameter values of an ideal Data Centre. Based on computation in Mat lab, they could draw a conclusion that the 3-tier architecture is the efficient one, however, with a high cost of implementation when compared to others. Further, it has been identified that if the unused network components were switched off, the power consumption could be reduced by a hefty 55%.

A more generalized approach has been showcased in [9]. It has been evaluated that the CISCO data center router Cisco Nexus X9536PQ: 36-port 40 Gigabit Ethernet QSFP + line card consumes 360W as typical operational power while its maximum power of operation is 400W.This indicates the need for energy management in such network devices for improved overall performance of the entire Data Centre.

Various models have been introduced for power consumption management. One of the simplest models is the Additive power model (Component-wise breakdown). Here, the power consumption is divided into two: static and dynamic. The energy consumption of a network device operating with a traffic load ρ can be expressed as:

$$E(\rho) = Estatic + Edynamic(\rho),$$

Where Estatic is the static power consumption independent from traffic and Edynamic(ρ) is the dynamic part that is a function of the traffic load ρ.Another model, proposed by Vishwanath et al. presented the power consumption P of an IP Router switch as the sum of the power consumed by its three major subsystems

$$P = Pctrl + Penv + Pdata,$$

Where the terms Pctrl, Penv, and Pdata represents the power consumption of the control plane, environmental units, and the data plane respectively. They further represented the Pctrl, Penv, and part of Pdata which are fixed as Pidle. The load dependent component of Pdata was expanded to two more terms based on packet processing energy and store & forward energy as,

$$P = Pidle + EpRpkt+ EsfRbyte,$$

WhereEp is the per-packet processing energy, and Esf is the per-byte store and forward energy which are constants for a given router/switch configuration. Rpkt is the input packet

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

rate and Rbyte is the input byte rate (Rpkt= Rbyte/L, where L is the packet length in bytes). Total energy used by a switch has been modelled in an additive power model.

2.3 Demand prediction strategies

In case of an individual CPU (static), a threshold is used to identify overloads. Static thresholds cannot be expected to adapt to changes in workload, as they are not suitable to handle system with unknown and dynamic workloads.

A possible approach to handle such cases is to perform decision-making based on statistical analysis of historical usage data. A server over-load forecasting technique based on time-series analysis of historical data is proposed in [10]. A dynamic server migration algorithm can predict the variable workload. In another approach, the periodic and interactive thresholds are considered by a trace-based workload migration controller. It balances the supply-demand ratio of the server, thereby minimizing the power consumption.

Dabbagh et al. [11] used k-means clustering and stochastic Weiner Filter for workload prediction. Here, the unused nodes were put to sleep, which helped reduce the power consumption.

Yet another approach is the Linear Regression based CPU Usage Prediction (LiRCUP) [12]. It utilizes linear regression to predict the utilization based on past CPU usage details (over an hour ago). This is used to manage overload and under-load situations in Data Centre in an efficient way.

One of the notable works was proposed by Prevost et al. [13]. Here neural network and auto-regressive linear prediction were used to forecast future demand profiles. The authors also concluded that a linear predictor model could produce more accurate functions.

2.6 Network traffic prediction strategies

Based on the Monte Carlo Tree Search, data center requests are processed in batches. The advantage with MCTS is that it can be used to perform a sampling-based look-ahead search. A tree structure is maintained, which is deepened in the direction which is most promising [14]. The edges and nodes represent actions and states respectively. Each node has a value κ. The κ is calculated using the trade-off between the cost of service and request blocking percentage. The higher the value of κ, the more preferred is the path. The algorithm considers both the path length and utilization. The advantage of this approach is that, time permitting, it may search for better path and DC assignments, thus more efficiently utilizing the network resources. Network operators may adjust the execution time of this approach based on the traffic load in order to improve network performance.

III. CONCLUSION

Data Centers, being one of the arenas with least energy efficiency, constitutes a crucial part for any IT Organization. With Machine Learning on the rise, it is high time to consider the inclusion of such technologies to improve the overall energy efficiency. The improvement of a single aspect of Data Centre may not yield the expected efficiency levels on the combined scale. It is advisable to implement a culmination of different solutions which caters to all the major group of equipment and their corresponding strategies for the overall improvement of PUE index.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

Acknowledgements

The authors would like to thank Dr.T. Sheela, HOD, Department of Information Technology, Sri Sairam Engineering College, Chennai, for her constant encouragement and support, which has greatly improved the manuscript.

IV. REFERENCES

[1]. Habibullah K.M., Rondeau E., Georges JP. (2018) Reducing Energy Consumption of Network Infrastructure Using Spectral Approach. In: Dastbaz M., Arabnia H., Akhgar B. (eds) Technology for Smart Futures. Springer, Cham.

[2]. Jun Liu, Optimizing the Energy Consumption of Servers and Networks in Cloud Data Centers. (unpublished).

[3]. Abbasi, et al., 2007; Amis, et al., 2007; Baker, 1981; Bandyopadhyay, 2003; Basagni, 1999; Lin, 1997;
Mellier, 2006; Chiasserini, et al., 2002.

[4]. Gao, "Machine Learning Applications for Data Center Optimization", Andrew Ng. "Neural Networks: Representation (Week 4)." Machine Learning. Retrieved from https://class.coursera.org/ml2012002. 2012. Lecture.

[5]. Z. Songa, X. Zhangb, C. Erikssona, Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Data Center Energy and Cost Saving Evaluation". In: The 7th International Conference on Applied Energy – ICAE2015.

[6]. J. Ni and X. Bai, "A review of air conditioning energy performance in data centers," Renewable and sustainable energy reviews, vol. 67.

[7]. Yuanlong Li, Yonggang Wen, Kyle Guan, and DachengTao, "Transforming Cooling Optimization for Green Data Center via Deep Reinforcement Learning".

[8]. MiyuruDayarathna, Yonggang Wen, Senior Member, IEEE, and Rui Fan, "Data Center Energy Consumption Modeling: A Survey".

[9]. N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing SLA violations", Proceedings of the 10th IFIP/IEEE Intl.Symp. on Integrated Network Management (IM), pp.119–128, 2007.

[10]. M. Dabbagh, B. Hamdaoui, M. Guizani, and A. Rayes, "Energy-efficient cloud resource management," in Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on. IEEE, 2014, pp. 386–391.

[11]. FahimehFarahnakian, PasiLiljeberg, and JuhaPlosila, "LiRCUP: Linear Regression based CPU Usage Prediction Algorithm for Live Migration of Virtual Machines in Data Centers", In: 2013 39th Euromicro Conference Series on Software Engineering and Advanced Applications

[12]. J. J. Prevost, K. Nagothu, B. Kelley, and M. Jamshidi, "Prediction of cloud data center networks loads using stochastic and neural models," in System of Systems Engineering (SoSE), 2011 6th International Conference on. IEEE, 2011, pp. 276–281.

[13]. MichałAibin, Krzysztof Walkowiak, Soroush Haeri and LjiljanaTrajkovic, "Traffic Prediction for Inter-Data CenterCross-Stratum Optimization Problems". In: 2018 International Conference on Computing,
Networking and Communications (ICNC): Optical and Grid Computing

[14]. Cazenave, "Nested Monte-Carlo search," in: IJCAI

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.03.241

InternationalJointConference on Artificial Intelligence2009