ISSN No:-2456-2165

Using a Deep Learning based Energy Consumption Reduction Model in Cloud Data Centres

Rufaro Sydney Madebwe
Department of Information Technology
Harare Institute of Technology
Harare, Zimbabwe

Tinashe Butsa
Lecturer; Department of Information Technology
Harare Institute of Technology
Harare, Zimbabwe

Abstract:- **Data centres have become a major part of computing, and with the advent of cloud computing their demand has soared. The increase in demand for cloud services has led to a data centre boom, however, the data centres tend to consume large amounts of power. The advent of Green computing has led to various researches into how to make computing on a large scale more sustainable. This has led to the evolution of power consumption prediction researches that are meant to help ease the use of power by data centres. In this regard, this research aims to look at ways to cope with the power consumption through adoption of Deep Learning to assist with feature selection. This method aims to look beyond the prior researches into power consumption which only looked at certain factors mainly consumption by the server and not the whole data centre. Key to this whole research area are the following phases: (i) performance monitoring and energy-related feature acquisition; (ii) essential feature selection; and (iii) model establishment and optimization.**

Keywords:- Green computing, Deep Learning, Feature Selection.

I. INTRODUCTION

Over the past two decades, there has been a notable increase in data centre scale. Application generated data in various industries is growing at an accelerated rate. With cloud computing taking on an increasingly important role in our lives, data centre energy management issues have become the focal point of many research endeavours[1][2]. Managing the energy issue of a data centre has become a major point of focus. Power Usage Effectiveness (PUE) is an industry metric used to measure power consumption. This is an effective engineering ratio which uses the following ratio:

$$PUE = \frac{\Sigma(Pm + Pe + PIT + Pother)}{P_{IT}} \quad (1)$$

Numerator shows the power consumption of all data centre facilities while denominator shows IT infrastructure energy consumption. As a measure of efficiency, a DC with a PUE ratio closer to 1 is highly energy efficient. Based on several industry surveys provided by Uptime Institute as well as other consulting institutes, the average power usage effectiveness (APUE) only improved from 2.5 in 2007 to 1.89 in 2011 then to 1.7 in 2014. An APUE of around 1.7 for enterprise data centres and between 1.5 and 1.6 for newly built data centres was still being recorded by the USA in 2017. Only a few DCs such as those of Facebook and Google were able to achieve PUE of 1.2 or lower.

However, energy efficiency improvements measured by PUE lately have been slowing down considerably. main reason has been that data centre efficiency optimisation becomes challenging to attain after a certain degree[3].

Energy efficiency, over the past two decades, has become modern cloud computing's major concern[4]. In light of this, many research works have been devoted to the provision of solutions that estimate power consumption of complex and dynamic datacentres. there has been employment of Machine Learning techniques for resource management and provision decisions. ML mainly focuses on building statistical models through training data, assisting in decision or prediction making about real world events. Some of the ML models include: neural networks(NNs), K-means, KNN, decision trees, and naive Bayes. in this research project, use of deep learning as a solution provider for the energy efficiency problem is proposed and looked at in detail. All things considered, data centres have various non-linear and complex devices, which can affect the accuracy of the experiment[5].

An overview of the remainder of the paper is as follows:

Section two mainly focuses on review of related works. Sections 3 and 4 detail the design and construction of the model. Section 5 looks at the comparison and analysis of experimental results. Summary of the paper and recommendations for future works are in Section 6.

II. RELATED WORKS

With energy efficiency in cloud computing becoming a hot theme in recent years, numerous solutions have been developed to model an aggregate power consumption for the server. This is due to the fact that servers provide the essential services in a datacentre so they are considered the datacentre's most power proportional components. Roy *et al.*[6] brought forward a server power model as a sum of CPU and memory consumption. The model was represented as:

$$E(A) = E_{cpu}(A) + E_{memory}(A) \quad (2)$$

Where cpu and memory are the particular components, while E is the energy consumption of the components when running algorithm A.

Tudor *et al*.[7] having considered more server components, came up with a power model defined as a function of CPU, memory and I/O device energy use. Song *et al*.[8] also described a similar power model through summation of CPU, memory, disk and NIC as follows:

$$E_{total} = E_{CPU} + E_{memory} + E_{disk} + E_{NIC} \quad (3)$$

Considering the server's key components and their resource utilisation levels, another energy model can be further constructed[9]

$$P_t = C_{CPU}U_{CPU} + C_{memory}U_{memory} + C_{disk}U_{disk} + C_{NIC}U_{NIC} \quad (4)$$

Where P_t is predicted power consumption at a time t, U_{cpu} is CPU utilisation, U_{memory} is memory access rate, U_{disk} is hard disk I/Orequest rate, and U_{nic} is the network I/Orequest rate; while C_{CPU}, C_{memory}, C_{disk} and C_{NIC} are CPU, memory, disk and NIC coefficients.

Furthermore, recent years have seen the performance of numerous studies dedicated to improving data centres' energy efficiency[10][11].

Zhou et al. [12] having employed different scenarios managed to show the power regression model's mathematical expressions in detail and it was verified that higher precision can be obtained through the use of the power regression model than any other regression models (linear, exponential and polynomial).

Given all these and a variety of other proposed works dealing with energy consumption, ensuring data centre energy efficiency has become a worthwhile endeavour. Application of machine learning as one of the innovative approaches is becoming more and more commonplace for cloud computing resource management and conservation problems.

Having detailed all the previous efforts, with great focus on their challenges and shortfalls, this paper looks to employ deep learning as a solution to the challenges data centres face. This was inspired by the aforementioned works with a goal to streamline energy efficiency.

III. METHODOLOGY

The methodology to be employed in this research comprises the following steps:

- Performance monitoring – with performance metrics as our primary data source, this step has the responsibility of monitoring all our metrics;
- Feature acquisition – energy related features are obtained at this stage;
- Feature selection – each feature's importance is calculated to establish its essentiality and therefore its selection as a critical feature;
- Modelling the energy consumption model – model is then built using Deep Learning;

Evaluation of the resulting experimental outcomes – results evaluation and demonstration of proposed model's effectiveness

- *Performance monitoring and feature acquisition*

In cloud computing, data acquisition is a key aspect of performance monitoring and analysis. To obtain server working status and timely collection of experimental data, server cluster performance metrics monitoring and management is extremely significant. Data monitoring and acquisition can be done through a variety of common performance monitors such as Nigel's performance Monitor (nmon), Ganglia, Zenoss, Cloudera Manager, Zabbix and Nagios. The major drawback associated with these is the excessive system overhead they cause.

Table 1 Energy related metrics

Metrics	Unit	Description
CPU utilization	percent	Percent of time CPU utilized
Memory use	gigabytes	Memory use for processes
Network I/O speed	Gigabytes/s	Data transmitted/received
Average load	Loads/procs	Average number of processes simultaneously in ready state during last minute
Number of M/R instructions	number	Instruction number of jobs
Disk utilisation	percent	Percent of time disk is utilised
Transmission and read/write ratio	percent	Data transfer and read/write ratio
Available space in file system	gigabytes	File system free space
Bytes consumed per CPU second	bytes	Average number of bytes consumed per CPU second
context switches rate	number	Number of switches between processes or threads
Power consumption	kw	Power consumed

- *Feature selection*

The monitoring and recording of any and all energy related features is implicit, however, different attributes have different weights when it comes to decision making. Data sets can have problems such as redundant and irrelevant attributes. To address the issue of information redundancy, we look at data dimensionality reduction algorithms.

First, we look at principal component analysis (PCA), a widely used unsupervised dimensionality reduction algorithm. It maximises sample variance to obtain the projection matrix after dimensionality reduction which helps to preserve as much as possible the global samples distribution properties. PCA tends to consume huge amounts of memory and high

ISSN No:-2456-2165

time complexity resulting in difficulties when calculating high dimensional data eigenvectors.

Linear discriminant analysis (LDA), is a supervised linear dimensionality reduction algorithm. Maximising the inter-class and intra-class dispersion quotient, LDA's strong discriminative ability can obtain a projection subspace by separating the different class samples as far as possible. LDA tends to ignore smaller inter-class dispersion with more focus being put on larger inter-class dispersion. This has the tendency of leading to fusion of different classes with minor inter-class dispersion which may lead to subset information loss and excessive reduction.

Locally linear embedding (LLE), dissimilar to PCA and LDA, is part of non-linear dimensionality reduction algorithms. LLE can represent global non linearity by leveraging local linearity all within the confines of maintaining original data properties unchanged. Subset features can be determined through mapping data from high dimensional to low dimensional space.

➢ *Energy consumption model*

For modelling and prediction of non-linear dynamic systems like cloud data centres, the deep learning approach is very effective.

➢ *Deep neural network architecture (DNNA)*

With deep learning, computational models comprising multiple processing layers for learning data representations with various levels of abstraction can be created.

DNNA, an energy consumption modelling framework has three main components as illustrated in Fig. 1:

- Input layer – contains *x* input nodes that are dependent on the input feature vector dimension and 1 bias node with a common value of 1
- Hidden layer – contains *a* layers with each layer having *z* hidden nodes and 1 bias node. Appropriate *a* and *z* values are dependent on comparison of multiple experiments.
- Output layer – contains only 1 output node representing the energy consumption.

When using DNNA, a set of vectors is specified. In this case we used a set of *x* vectors i.e. I_1, I_2, … I_x. Electrical energy consumption denotes the DNNA estimated result at the output layer. Using DNNA in forecasting, an input matrix with *c* feature vectors gets multiplied by matrix model parameters thereby producing the first hidden state matrix. Subsequent hidden layers forward propagate intermediate results based on the same computational process until a predicted energy consumption value is calculated.

For deep learning, two approaches were adopted. The first was Graphical User Interface (GUI) based and the second was a more hands on approach. For the GUI based approach, Waikato Environment for Knowledge Analysis (WEKA) was used. It offered a wide variety of training and testing methods for machine learning. A WEKA deep learning plugin is also available but mainly for Java. The main challenge with WEKA was when model construction was supposed to take place. That is where the second approach came in handy. With the second approach, Python was the preferred route. Python, while already having deep learning capabilities, can also allow the development and training of a model all in one program instance.

The test environment was software based. In an attempt to recreate a data centre locally, CloudSimPlus[13] was used. This is a Java based simulation environment that mimics nearly all data centre resources. It is however resource intensive so minimal use and resource allocation was of the essence.

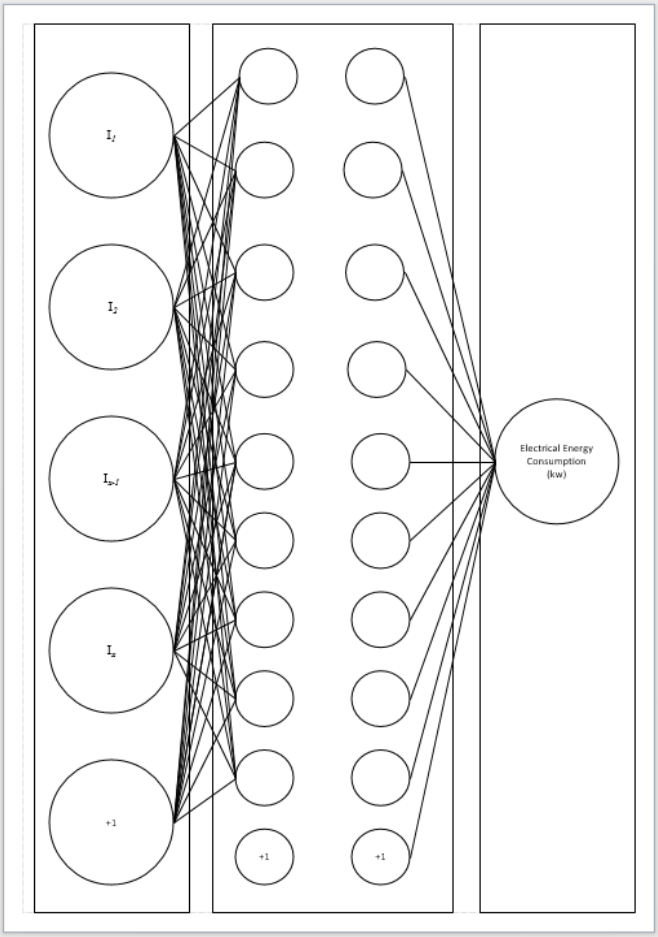

Fig. 1 DNNA

IV. IMPLEMENTATION AND RESULTS

For our model, the Hadoop framework will be substituted to simulate a cloud environment. Two virtual machines will be set up to run Hadoop MapReduce jobs and the performance metrics from the running jobs are to be collected. These data can then used to train and configure the energy consumption model. Given the nature of MapReduce jobs, they fall into one of three categories: compute-intensive, I/O intensive and transactional web applications.

For an exhaustive result, the Hadoop cluster is to be subjected to various assignments. This allows proper monitoring of performance and energy usage. Under different circumstances, even the same feature can have different contributions to energy consumption in different domains.

Disk traffic is of more importance in I/O intensive applications whereas it may not be crucial for compute-intensive jobs.

The experimental environment setup will be as such, VirtualBox installed and two Windows 10 virtual machines setup and configured. These would act as our simulated data centre. The Hadoop Distributed File System is to then be setup to run with one being the namenode and the other the datanode. From this setup certain key metrics will be extracted that can be used for the testing and evaluation of data centre power consumption.

V. CONCLUSION

The preceding project report was done with special consideration to the fact that data centres, being an integral part of our day to day lives, are also proving to be less energy efficient. With the advent of Green Computing and in the wake of environmental awareness, data centres will have to find ways of managing their workloads to reduce their energy use. With that in mind, this research project was embarked on to see if its possible to get a better use of data centre energy while maintaining performance. The use of a deep learning based model makes it possible to maintain performance while keeping energy consumption levels within a manageable range. Given this discovery, it is possible for data centres to maximise on performance with considerable reduction in energy usage.

REFERENCES

[1]. J. Shuja, K. Bilal, S. A. Madani, and M. Othman, "Survey of Techniques and Architectures for Designing Energy-Efficient Data Centers," no. October 2017, 2016, doi: 10.1109/JSYST.2014.2315823.
[2]. M. Dayarathna, Y. Wen, S. Member, and R. Fan, "Data Center Energy Consumption Modeling : A Survey," vol. 18, no. 1, pp. 732–794, 2016.
[3]. S. Hasan, F. Alvares, T. Ledoux, and J. L. Pazat, "Investigating Energy consumption and Performance trade-off for Interactive Cloud Application," no. June, 2017, doi: 10.1109/TSUSC.2017.2714959.
[4]. F. Jalali, K. Hinton, R. Ayre, T. Alpcan, and S. Member, "Fog Computing May Help to Save Energy in Cloud Computing," no. March 2018, 2016, doi: 10.1109/JSAC.2016.2545559.
[5]. H. A. Kurdi and S. M. Alismail, "LACE : A Locust-Inspired Scheduling Algorithm to Reduce Energy Consumption in Cloud Datacenters," *IEEE Access*, vol. 6, pp. 35435–35448, 2018, doi: 10.1109/ACCESS.2018.2839028.
[6]. S. Roy, "An Energy Complexity Model for Algorithms," 2013.
[7]. B. M. Tudor and Y. M. Teo, "On Understanding the Energy Consumption of ARM-based Multicore Servers," pp. 267–278, 2013.
[8]. S. L. Song, K. Barker, and D. Kerbyson, "Unified Performance and Power Modeling of Scientific Workloads," 2013.
[9]. I. Alan, "Energy-Aware Data Transfer Algorithms."
[10]. A. Hameed, R. Ranjan, and P. P. Jayaraman, "A survey and taxonomy on energy efficient resource allocation techniques for cloud computing systems A survey and taxonomy on energy efficient resource," no. June, 2014, doi: 10.1007/s00607-014-0407-8.
[11]. Z. Li *et al.*, "A Survey on Modeling Energy Consumption of Cloud Applications : Deconstruction , State of the Art, and Trade-off debates," no. August, 2017, doi: 10.1109/TSUSC.2017.2722822.
[12]. Z. Zhou, J. H. Abawajy, F. Li, Z. Hu, and M. U. Chowdhury, "Fine-Grained Energy Consumption Model of Servers Based on Task Characteristics in Cloud Data Center," 2018.
[13]. M. M. Freire, "CloudSim Plus."