International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

A Survey on Neural Network Based Minimization of Data Center in Power Consumption

E. Brindha[1], S. Sobitha Ahila[2]
[1, 2]Easwari Engineering College, Tamilnadu, India
Email address: [1]brindha235@gmail.com, [2]sobithaahila@gmail.com

***Abstract*—** Cloud computing is the current technology used for sharing and accessing resources via internet. Reducing power consumption is an essential requirement for cloud resource providers to decrease operating costs. We employ the predictor to predict future load demand based on historical demand. According to the Prediction, the algorithm turns off unused servers and restarts them to minimize the number of running servers. Power consumption can be regulated by using proper load balancing technique. Load balancing is done so to distributing the load fairly amidst the servers and also a scheduling technique is followed to selectively hibernate the servers to optimize the energy consumption. The load balancing is based on load prediction and server selection policy. Neural Network is used for load prediction, which predicts the future load based on past historical data. The servers can be monitored and given ranking based on their reliability record and this information is used as a criterion while performing load balancing.

***Keywords*—** Cloud computing; eucalyptus; load balancing; neural network.

I. INTRODUCTION

Cloud computing is a recent advancement where in the Information Technology (IT) infrastructure and applications are provided as "services" to end-users under a usage-based payment model. The cloud makes it possible to access the information from anywhere at any time. While a traditional computer setup requires the user to be in the same location where the data storage is present, the cloud makes the storage and retrieval location independent.

Cloud computing offers delivery of on-demand computing resources everything from application to data centre resources over the Internet. It offers guaranteed services. It allows customers to scale up and down their resources based on the dynamic needs.

One of the most pressing issues in cloud computing is the resource management. Resource management problems include allocation, provisioning, requirement mapping, adaptation, discovery, brokering, estimation and modeling. Resource management in cloud computing offers the benefits like scalability, quality of service, optimal utility, reduced overheads, improved throughput, reduced latency, specialized environment, cost effectiveness and simplified interface. This research focuses on some of the important resource management techniques such as resource provisioning, resource allocation, resource mapping and resource adaptation. It brings out an exhaustive survey of such techniques in cloud computing and also put forth the open challenges for further research.

A. Cloud Service Models

The services offered by the Cloud service provider are generally classified into three types of service models. Software as Services (SaaS):- End-user application is delivered as a service. Platform and infrastructure is abstracted, and can be deployed and managed with less effort. Cloud-based applications run on remote servers in the cloud that are owned and operated by others and that connect to users' computers via the Internet and a web browser. Platform as a Services: - Services is provided as application platform onto which custom applications and services can be deployed. PaaS services can be built and deployed more inexpensively, although services need to be supported and managed. Platform as a Service provides a cloud-based environment with everything required to support the complete life cycle of building and delivering web-based cloud applications, without the cost and complexity of buying and managing the underlying hardware, software, provisioning and hosting. With PaaS one can develop applications and get to market faster, deploy new web applications to the cloud in minutes and reduce complexity with middleware. Infrastructure as a Services (IaaS):- Consumers control and manage the systems in terms of the processors, memory, and storage and network connectivity but do not have control over the cloud infrastructure. Infrastructure as a Service provides companies with computing resources on a pay-per-use basis.

B. Deployment Models

Cloud computing offers three types of deployment models, namely, Private cloud:-The cloud infrastructure is deployed and maintained for a specific organization. The datacenter may be in-house or on a third party premises. Public cloud:- The cloud infrastructure is available to the public on a commercial basis by a cloud service provider. This enables a consumer to develop and deploy a service in the cloud with little financial outlay compared to the capital expenditure requirements normally associated with other deployment options. Hybrid cloud: - The cloud infrastructure consists of a number of clouds of any type, but the clouds have the ability through their interfaces to allow data and/or applications to be moved from one cloud deployment model to another. This can be a combination of private and public clouds that support the requirement to retain some data within an organization and also the need to offer services to the general public.

E. Brindha and S. Sobitha Ahila, "A survey on neural network based minimization of data center in power consumption," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 127-130, 2016.

C. Neural Network

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions. Other advantages include: Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage. An artificial neuron network is a data processing system consisting large number of highly interconnected processing elements as artificial neuron in a network structure. By using cloud computing service for neural networks we can design any type of neural networks.

D. Load Analysis and Prediction

Load analysis is the process of analyzing the load of the server in the datacenter on time period basis. The load analysis will be done more than once to provide a better result.

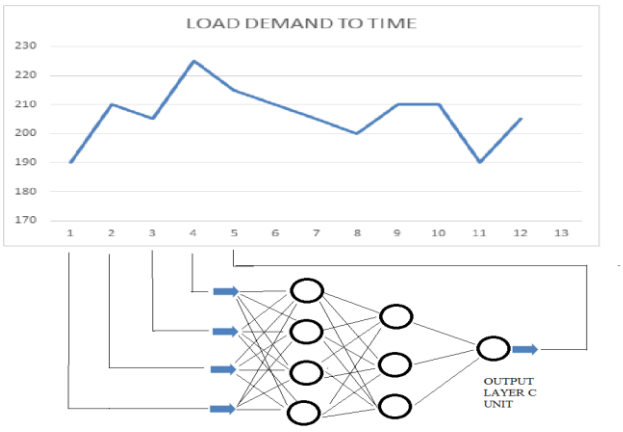

Fig. 1. A three layer neural network prediction.

Load prediction will be carried out based on the analyzed data using neural network predictor. The neural network perceptrons are applied to predict the future load of the applications. The perceptrons are trained in a supervised manner with a back propagation algorithm. Perceptrons uses the values: Number of external inputs, one internal input. Threshold and single output are prediction. The figure 1 as three layer neural network predictor in operation with a time series input is plotted. Load prediction is necessary for improving power grid management. In recent years, several methodologies have been advanced to accurately predict power load. However, most of them cannot handle the prediction well during unanticipated situations, which may escalate into serious power failures. We introduce a methodology seeking improvements in power load analysis and prediction for unanticipated situations. The new methodology exploits the synergisms found between artificial neural networks and fuzzy logic to detect load transients in a quickest manner. It subsequently identifies the transients and provides on line prediction for abnormal conditions.

When the network is run, each layer performs the calculation on the input and transfers the result y_{n+1} to the next layer.

$$\mathbf{Y_{n+1}=h((}\sum_{i=1}^{n}(X_i + \mathbf{b}))/n) \quad (1)$$

The above equation (1) is used for prediction of future load based on the input value. Where Y_{n+1} provides the output of the current node and n is the number of nodes in the previous layer, x_i is the input of the current node from the previous layer, b is the bias value and w_i is the modified weight based on the mean square error. Here the neural predictor is developed and the experiment is performed to prove its highly accurate prediction ability which fits in dynamic real time.

II. Literature Survey

A. Power-Aware Scheduling of Virtual Machines in DVFS-Enabled Clusters

Gregor von Laszewski et al 2012, Scheduling virtual machine in a compute cluster to reduce power consumption via the technique of dynamic voltage frequency scaling. Specifically, we design of an efficient scheduling algorithm to allocate virtual machine in DBFS-enabled cluster. A computer cluster provides various virtual machines and when a job arrives at the cluster, the cluster scheduler allocates the job with preconfigured virtual machine then starts it on proper compute nodes. The job is executed in the virtual machine. After the job is executed, the virtual machine is shut-down.

Cloud computing technologies, the needs for efficient algorithm to minimize wasted server energy becomes increasingly important. As such, the field of green computing provides a way to prevent unnecessary emission from contributing to global warming to save large amount of money on operational cost.

B. Performance Evaluation of a Green Scheduling Algorithm for Energy Saving in Cloud Computing

Truong vinh Truong Duy et al 2010, we employ the prediction to predict future load demand based on historical demand. According to the prediction, the algorithm turns off unused servers, thus minimizing the energy use at the points of consumption to benefits all other levels. For evolution, we simulation with two load traces. The result show that the P20 mode can save up to 46.3% of power consumption with a drop rate of 0.03% on one load trace, and a drop rate of 0.12% with a power reduction rate of 46.7% on the other.

E. Brindha and S. Sobitha Ahila, "A survey on neural network based minimization of data center in power consumption," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 127-130, 2016.

Green scheduling Algorithm which makes use of neural network based predictor for energy saving in cloud computing. The prediction is exploited to predict future load demand based on collected historical demand. The algorithm used the prediction in making turning off/on decision to minimize the number of running servers.

C. *A Study on Server Sleep State Transition to Reduce Power Consumption in a Virtualized Server Cluster Environment*

V. K. Mohan Raj et al 2012, Reducing power consumption is an essential requirement for cloud resources providers to decrease operating costs. One of the options to reduce power consumption is to reduce the number of servers in IDLE (unused) state as these IDLE state can be reduced by turning off IDLE servers or transitioning these IDLE servers to low power SLEEP state. With virtualization being the backbone to provision cloud computing services, we use simulation to study and report the impact of using SLEEP state on the server with its virtual machine servicing application workload requests. The average response time per request, our simulation result are using SLEEP state server level we achieve a 2% saving in average power usage and around 27% saving in average response time per request.

The work deal with physical SLEEP state transition impact to overall power consumption and response times in a physical server environment. In this work heuristic approaching for SLEEP state at server level and using VM"s to processing application requests the virtualized environment. To show saving to overall average per application request response time and marginal workload specific power consumption saving.

D. *The Eucalyptus Open Source Cloud-Computing System*

D. Nurmi et al 200, Cloud computing that implement what is commonly referred to as infrastructure as service systems that give users the ability to run and control entire virtual machine instances deployed across a variety physical resources. The operational aspects of the system and architectural trade-off that we have made in order to allow the eucalyptus to be portable, modular and simple to use on infrastructure commonly found within academic settings, we provides evidence the eucalyptus enables users familiar with existing grid and HPC systems to explore new cloud computing functionality while maintaining access to existing, familiar application development software and grid middle ware.

Eucalyptus system built to allow administrators and researchers the ability to deploy an infrastructure for user-controlled virtual machine creation and control a top existing resources. It hierarchical design targets resource commonly found within academic and setting including but not limited to small and medium Linux clusters, workstation pools, and server farms. We use a virtual networking solution that provides VM isolation, high performance and a view of network that is simple and flat.

E. *Genetic Load and Time Prediction Technique for Dynamic Load Balancing in Grid Computing*

Z. Akhtar et al 2007, a genetic new algorithm based task scheduling has been proposed and its performance has been proved on many applications on the grid. This addresses the problem with Priority based Load balancing in cloud for Data intensive applications. The requests are classified based on a certain parameter like source IP address, bandwidth of the cloud consumer, affordability of consumer, SLA required etc. Normally the data is retrieved from a distributed database environment since it provides high availability and redundancy. The scheduling of resources has been classified from three Cloud Service Model perspectives. The first model is based on Software as a Service (SaaS) perspective which aims at providing efficient resource allocation using cost optimization and priority models. Cost optimization is done by consolidating the hardware and software resources required to run the datacenters on specified SLAs and also aim to provide maximum profit to the cloud service provider. In a datacenter during the period when resources are available in excess, the resources can be auctioned. A bidding policy is defined to ensure a fair and optimum profit generated for the cloud service provider. Auctioning is generally the technique used to increase the price of an object in which multiple bids are compared to determine the highest bid and the object.

F. *Load Balancing Incoming IP Request Across a Farm of Clustered my SQL Server*

M. Kaitsa, et al 2009, MySQL Cluster Database provides services with full capability to cover the peak demands. In Cloud environment based on demand, requests are assigned to the different VM's. Requests are sent across a farm of MySQL cluster servers in which the data is replicated to avoid single point failure and also does load balancing, there-by a better response time is achieved. Advanced eager scheduling achieves fault tolerance and load balancing by dynamically breaking down the tasks and by performing parallel computing.

The last perspective model is based on Infrastructure which aims at energy saving and power-aware cloud based on past data and current data. A green scheduling algorithm with neural based prediction technique, predicts the future load based on the past history. However this approach causes heavy overhead due to stop/start of the servers dynamically and hence leads to strong performance degradation. This thesis provides a power optimized solution based on neural network trained with back propagation algorithm. The future load is predicted using the trained set which has less mean square error and also high reliability.

III. CONCLUSION

This growing crisis in power shortage has brought a concern in existing and towards reliable optimal load balancing with the help of load prediction using neural network model. The sample loads are gathered from a datacenter and used an input for prediction. Using past input data the future load is predicated using the trained data set with less mean square error and high reliability. From the predication value the required number of server with high

E. Brindha and S. Sobitha Ahila, "A survey on neural network based minimization of data center in power consumption," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 127-130, 2016.

ISSN: 2454-1532

reliability is selected to perform load balancing to optimize power consumption with higher reliability.

REFERENCES

[1] Z. Akhtar, "Genetic load and time prediction technique for dynamic load balancing in grid computing," *Infomation Technology Journal*, vol. 6, pp. 978-986, 2007.

[2] V. K. M. Raj, "A study on server sleep state transition to reduce power consumption in a virtualized server cluster environment," *IEEE Fourth International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1-6, 2012.

[3] D. Nurmi, "The eucalyptus open sources cloud computing system," Department of Computer Science, University of California, 2000.

[4] G. V. Lazewski, "Power aware scheduling of virtual machines in DVFS enables clusters," Pervasive Technology Institute, Indiana University, Boomington, 2012.

[5] T. V. T. Duy, Y. Sato, and Y. Inoguchi, "Performance evaluation of a green scheduling algorithm for energy saving in cloud computing," *IEEE International Sysmposim on Parallel and Distributed Processing, Workshops and Phd Forum (IPDPSW)*, pp. 1-8, 2010.

[6] M. Kaitsa, I. Stavrakas, T. Kontogiannis, I. Daradimos, M. Panaousis, and D. Triantis, "Load balancing incoming IP request across a farm of clustered MySQL servers," *Proceeding of the International Conference on Computer as a Tool*, Warsaw, pp. 546-550, 2009.

E. Brindha and S. Sobitha Ahila, "A survey on neural network based minimization of data center in power consumption," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 127-130, 2016.