JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Thermal Prediction using Machine Learning for Efficient Energy Management of Cloud

Prof. Namrata Ghuse[1], Sanjana Desale[2], Abhishek Doltade [3], Ravina Bachhav[4], Sunayana Guthale[5]

[1]Assistant Professor [2 3 4 5]Student
[1 2 3 4 5]Department of Computer Engineering
[1 2 3 4 5]Sandip Institute of Technology and Research Center, Nashik, India

---***---

Abstract - *Thermal prediction for efficient energy management of clouds using machine learning involves using predictive models to optimize the cooling and energy consumption of cloud data centers. Cloud data centers require substantial cooling to maintain optimal operating temperatures for servers and other hardware. By accurately predicting the thermal behavior of the data center, energy consumption can be optimized, resulting in cost savings and improved overall efficiency. Temperature estimation is a non-trivial problem due to thermal variations in the data center. Existing solutions for temperature estimation are inefficient due to their computational complexity and lack of accurate prediction. Energy-efficient Cloud Infrastructure Allocation Of resources Framework is gaining recognition because it focuses on cloud data management to increase profit while minimizing costs. For efficient resource management, accurate host temperature prediction is essential. In this paper we proposed a method for prediction of thermal energy and scheduling it properly.*

Keywords: Cloud computing, machine learning, Resource Management System (RMS), Reinforcement Learning, Linear Regression, Green Cloud, Gradient Boosting (XGBoost).

1. INTRODUCTION

Driven by more's law the trend in increasing performance of CPUs has seen as collateral effects the rapid increase of power consumption and power density that are at the root of performance degradation, acceleration of chip ageing, and cooling costs. Cooling and heat management are rapidly becoming the key limiters for high performance processors, especially for HPC and data centers which typically host clusters of hundreds (and sometimes even thousands) of high-performance processors. Thus, approximately 50 percent of the energy consumed by data centers is used for powering the cooling infrastructure. The remaining energy is used for computation and causes the temperature ramp-up. This trend is globally visible, in 2009 about 2 percent of the global electricity production was consumed to operate data centers worldwide, and accounted for an estimated 0.7 percent of the global energy-related CO2 emissions[1].

2. The research[2,3] community and leading electronics companies have invested significant efforts in developing thermal control solutions for computing platforms, limiting the overhead imposed by worst case thermal design in both cost and performance. Indeed, even if design optimization can improve the architectural thermal efficiency, on-chip silicon performance and usage show highly spatial and temporal variability; this reflects in the inefficiency of static approaches to solve thermal issues, causing HW damage, reliability loss and cooling costs overhead.

3.

4. Modern cloud data center's rack-mounted servers can consume up to 1,000 watts of power each and attain peak temperature as high as 100 C [4]. The power consumed by a host is dissipated as heat to the ambient environment, and the cooling system is equipped to remove this heat and keep the host's temperature below the threshold. Increased host temperature is a bottleneck for the normal operation of a data center as it escalates the cooling cost. It also creates hot spots that severely affect the reliability of the system due to cascading failures caused by silicon component damage. The report from Uptime Institute [5] shows that the failure rate of equipment doubles for every 10 C increase above 21 C. Hence, thermal management becomes a crucial process inside the data center Resource Management System (RMS).

5.

6. To minimize the risk of peak temperature repercussions, and reduce a significant amount of energy consumption,

ideally, we need accurate predictions of thermal dissipation and power consumption of hosts based on workload level. In addition, a scheduler that efficiently schedules the workloads with these predictions using certain scheduling policies. However, accurate prediction of a host temperature in a steady-state data center is a non-trivial problem. This is extremely challenging due to complex and discrepant thermal behavior associated with computing and cooling systems. However, predicting future temperature based on the change in workload level is equally necessary for critically important RMS tasks such as resource provisioning, scheduling, and setting the cooling system parameters.

A system model for predictive thermal management consists of three modules i.e. training, process and testing. In training module the dataset is given as input and the system is trained by processing. After that the system is tested by giving inputs from the dataset.

2. LITERATURE SURVEY

Existing temperature prediction methods are erroneous, difficult, or computationally expensive. The frequently used theoretical analytical models [6], [7], [8], [9], [10] based on mathematical relationships between diverse cyber-physical components lack scalability and precise temperature prediction. Furthermore, theoretical models fail to account for various variables that influence temperature behaviour, and they must be adjusted for different data centres. Computational Fluid Dynamics (CFD) models are also commonly employed for accurate forecasts [11], [12], but their great complexity necessitates a huge number of computation cycles. Building and running these CFD models might take hours or days, depending on the complexity of the data centre [13]. CFD models are beneficial for initial design and calibration of data centre layout and cooling settings, but they are impractical for realtime operations that are dynamic and demand immediate online decisions (e.g., scheduling in large scale clouds). Furthermore, CFD simulation involves both computational (e.g., Data Centre layout, open tiles) and physical factors, and changes to these parameters necessitate costly model retraining [14].

In 2018 Emmanuel N. et al. set out to create a strategy for the implementation of a computing infrastructure for communication and information technology centers (ICTs) in tertiary institutions in Nigeria. According to recent research, cloud computing will become the norm in technological advances and will be extremely beneficial to organizations. ICT units are found in all educational institutions, and they are in charge of providing ICT systems and facilities for administrative, educational, study, and teacher education as a whole.

In 2017 Arti Singh et al.introduced a new agent-based automated system structure algorithm that includes demand computation and computerized network design stages and not only searches for integrated solutions but also recognizes and lowers the cost of virtual machines that are only used by on￾demand offerings. Several power management concerns have been stated by Samah Ibrahim Alshathri (2016) in his categorization of cloud computing systems. In addition, virtualization, migrations, and work system architectures were studied to reduce power usage in cloud data centers. The use of a novel management concept will aid in the design and monitoring of the matching processing times between data centers and inbound jobs.

In 2016 J. Ye, Y. He, X. Ge, \& M. Chen in proposed a traffic engineering method for Wireless Mesh Networks using joint selection of link-channel pairs at each forwarding router. They used the method for the allocation of required powers in order to boost up overall network throughput. Chia-Ming Wu et al. (2014) suggested a cloud datacenter planning problem within the parameters of the proposed service scaling technique. The suggested programming method can effectively increase the utilization of resources, resulting in lower energy consumption when jobs are executed. According to the results obtained, the method can reduce energy usage more than other schemes.

In 2013 Z. Xiao et al. proposed a resource management scheme that make use of changing demand in order to adaptively multiplex virtual and physical resources. The concept of “skewness” was introduced to measure the unevenness in the multi￾dimensional resource utilization of a server. By minimizing skewness, different types of workloads were combined nicely and improved the overall utilization of server resources. They developed a set of heuristics that prevented overload in the system effectively while saving energy used.

In 2011 J. Baliga et al. studied the problem energy consumption in in public and private cloud systems, in particular, within data processing and data storage units. The analysis considered both public and private clouds. The authors measured energy consumption in three different cloud services: processing, software, and storage as a service. They concluded that public cloud consumes around four times more power than private. They also showed that cloud based systems consume more energy than traditional computing paradigms even while adopting optimization methods.

In 2011 I. S. Moreno, and J. Xu discussed the importance of energy savings without degrading the performance in cloud computing which represented a business model where the satisfaction of customers has high priority. The concept for performance preservation introducing policies and evaluations were proposed in their methodologies. Some additional variables were introduced such as workload and hardware heterogeneity, workload networking and server's optimal utilization. The authors stated that cloud systems still have room for energy optimization while maintaining the level of service provided to users.

In 2011 Q. Chen, P. Grosso, K. v. d. Veldt, The GreenClouds project in the Netherlands investigates a system-level approach towards Greening High-Performance Computing (HPC) infrastructures and clouds. Q. Chen et al. built a linear power model that represents the behavior of a single work node and includes the contribution from individual components, i.e. CPU, memory and HDD, to the total power consumption of a single work node. Three power metrics, i.e. power, power efficiency and energy, were considered in profiling virtual machines with respect to different high performance computing workloads. Q. Zhang et al. (2011) presented a method for resource management that can dynamically adjust both supply and price in order to maximize the provider's revenue and customer satisfactions in terms of VM scheduling delay. They applied their method on a single provider scenario and conducted their method evaluation using Amazon EC2. The mechanism was based on a constrained discrete-time finite-horizon optimal control formulation and used Model Predictive Control (MPC) to find its solution for designing their dynamic algorithm.

In 2010 A. Beloglazov et al. proposed and evaluated efficient resource management policy for virtualized cloud data centers. They presented a heuristics based method to minimize energy consumption in VMs allocation using QoS as constraints. They showed that switching off the idle servers by making use of dynamic reallocation of VMs can reduce energy consumption in cloud data centers. R. Buyya et al. [11] in 2010 presented a state of the art review on energy-efficient computing. The authors suggested that cloud systems require new architectural concepts to keep up with energy efficiency requirements. In resource allocation methodologies that maintain quality of service should be addressed, in particular, while addressing power usage characteristics of the devices. The authors then suggested a resource allocation algorithms that is based on the dynamic characteristics of virtual machines

3. SYSTEM METHODOLOGY

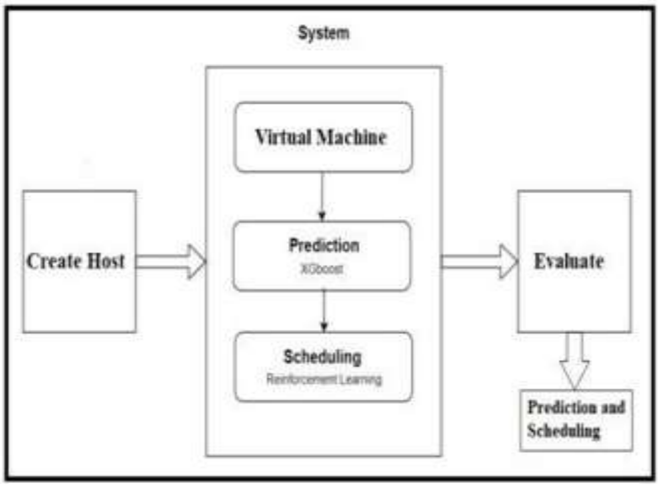

Fig. System Architechture

The proposed system first creates the virtual machines and then the hosts. After that the prediction and allocation is done. The choice of regression-based algorithms for our problem is natural since we aim to estimate the numerical output variable i.e., temperature. In the search for suitable prediction mechanisms, we have explored ML algorithm regression techniques, such as Linear Regression (LR) and an ensemble learning technique called gradient boosting, specifically, eXtreme Gradient Boosting (XGBoost). Linear regression can be used for thermal prediction by establishing a relationship between the independent variables (e.g., environmental factors, system parameters) and the dependent variable (e.g., temperature). The model assumes a linear relationship between the variables and aims to find the best-fitting line that represents this relationship. XGBoost (eXtreme Gradient Boosting) is a powerful machine learning algorithm that can be effectively applied to thermal prediction tasks. It is an ensemble method that combines the predictions of multiple weak learners (decision trees) to create a strong predictive model.

We predict the host ambient temperature (Tamb) which is a combination of inlet temperature and CPU temperature [20]. The reason to consider ambient temperature instead of CPU temperature is manifold. First, by combining the inlet and CPU emperature, it is feasible to capture thermal variations that are induced by both the inlet and CPU temperature . Second, at a data center level, cooling settings knobs are adjusted based on host ambient temperature rather than individual CPU temperature.

4. RESULT

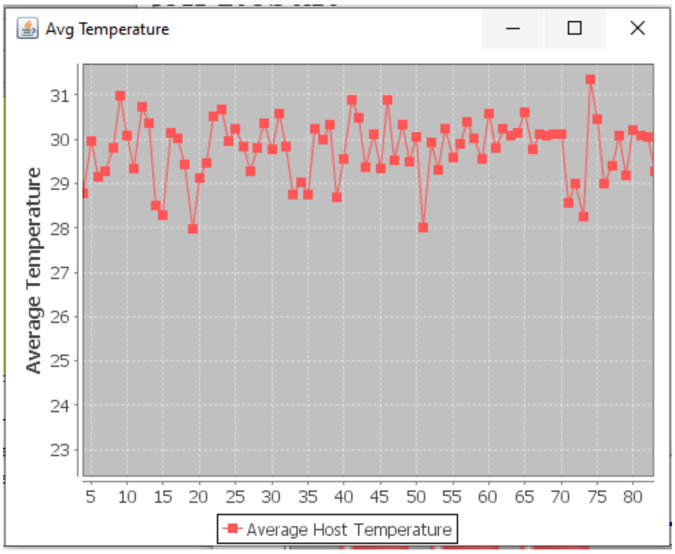

Figure1 : Average host temperature

From above figure we can see that the host workload is manage so well so the average temperature is maintained properly.

Comparative analysis of algorithm is shown in figure below.

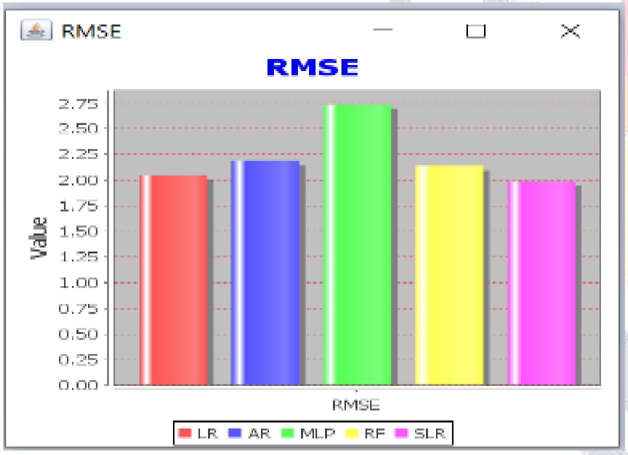

Figure 2: Comparative Analysis

5. CONCLUSIONS

Estimating the temperature in the data center is acomplex and non-trivial problem. Existing approaches for temperature prediction are less-accurate and computationally expensive.

Optimal thermal management with accurate temperature prediction canreduce the operational cost of a data center and increase reliability. We present our proposed system in this paper and then we discussed the results also.

6. REFERENCES

[1] M. Alhamad, T. Dillon, E. Chang, "Conceptual SLA Framework for Cloud Computing", Proceedings of the 2010 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST), 2010, p. 606 -610.

[2] A. Beloglazov, and R. Buyya,"Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", Journal of Concurrency and Computation: Practice & Experience archive, Volume 24, Issue 13, September 2012, p. 1397-1420.

[3] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. D. Rose and R. Buyya, "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms", Journal of Software: Practice and Experience, Volume 41, Issue 1, 2011, p. 23-50.

[4] G. Koutitas and P. Demestichas, "Challenges for Energy Efficiency in Local and Regional Data Centers", Journal of Green Engineering, ISSN: 1904-4720, River Publisher, October 2010, p. 1-32.

[5] K. S. Park, and V. S. Pai, "CoMon: a mostly-scalable monitoring system for PlanetLab" , In the Association for Computing Machinery Special Interest Group on Operating Systems (ACM SIGOPS) Review, Volume 40 Issue 1, January 2006, p. 65-74.

[6] A. Rawson, J. Pfleuger, T. Cader, "Data Center Power Efficiency Metrics: PUE and DCIE", 2008.

[7] J. P. D. Comput, A. Iulian, F. Pop, and I. Raicu, "New scheduling approach using reinforcement learning for heterogeneous distrib□uted systems," J. Parallel Distrib. Comput., vol. 117, pp. 292–302, 2018

[8] Aransay, M. Z. Sancho, P. A. Garc□ıa, and J. M. M. Fern□andez, "Self-organizing maps for detecting abnormal thermal behavior in data centers," in Proc. 8th IEEE Int. Conf. Cloud Comput., 2015, pp. 138–145

[9] M. A. Adnan, R. Sugihara and R. Gupta, "Energy Efficient Geographical Load Balancing via Dynamic Deferral of Workload ", Proceedings of the 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), June 2012, p. 188-195.

[10] Q. Chen, P. Grosso, K. v. d. Veldt, C. de Laat, R. Hofman, and H. Bal, "Profiling Energy Consumption of VMs for Green Cloud Computing", Proceedings of the 2011 IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing, DASC '11, IEEE Computer Society Washington, DC, USA 2011, p. 768-775 14

[11] A. Beloglazov, and R. Buyya, "Energy Efficient Resource Management in Virtualized Cloud Data Centers", Proceedings of the 2010 10th IEEE/ACM International Conference on

Cluster, Cloud and Grid Computing, IEEE Computer Society Washington, DC, USA, 2010, p. 826-831.

[12] K. Zhang et al., “Machine learning-based temperature prediction for runtime thermal management across system components,” IEEE Trans. Parallel Distrib. Syst., vol. 29, no. 2, pp. 405–419, Feb. 2018.

[13] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos, “Energy-efficient thermal-aware task scheduling for homogeneous high-perfor mance computin{g data centers: A cyber-physical approach,” IEEE Trans. Parallel Distrib. Syst., vol. 19, no. 11, pp. 1458–1472, Nov. 2008.

[14] H. Sun, P. Stolf, and J.-M. Pierson, “Spatio-temporal thermal aware scheduling for homogeneous high-performance computing datacenters,” Future Gener. Comput. Syst., vol. 71, pp. 157–170, 2017.

[15] S. Zhang and K. S. Chatha, “Approximation algorithm for the temperature aware scheduling problem,” in Proc. Int. Conf. Com put.-Aided Des., 2007, pp. 281–288.

[16] S. Ilager, K. Ramamohanarao, and R. Buyya, “ETAS: Energy and thermal-aware dynamic virtual machine consolidation in cloud data center with proactive hotspot mitigation,” Concurrency Com put., Practice Experience, vol. 31, no. 17, 2019, Art. no. e5221.

[17] J. Gao, “Machine learning applications for data center opti⍰mization,” Google White Paper, 2014.

[18] M. Cheng, J. Li, and S. Nazarian, “DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers,” in Proc. Asia South Pacific Des. Autom. Conf., 2018, pp. 129–134.

[19] C. Imes, S. Hofmeyr, and H. Hoffmann, “Energy-efficient applica⍰tion resource scheduling using machine learning classifiers,” in Proc. 47th Int. Conf. Parallel Process., 2018, pp. 45:1–45:11

[20] R. Bianchini et al., “Toward ML-centric cloud platforms,” Com⍰mun. ACM, vol. 63, no. 2, pp. 50–59, 2020.