Literature Survey of Energy Efficient Scheduling Algorithm for Cloud Computing Environment

Sheetal Aggarwal[1], Vandana Dabas[2]

[1]Student, [2]H.O.D Computer Science & Engg. Deptt.,
SET, Ganga Technical Campus, Soldha

Abstract- **Cloud services are typically made available and usable via a private cloud, community cloud, public cloud or hybrid cloud. Generally speaking, services provided by a public cloud are offered over the Internet and are owned and operated by a cloud provider. Some examples include services aimed at the general public, such as online photo storage services, e-mail services, or social networking sites. However, services for enterprises can also be offered in a public cloud. In a private cloud, the cloud infrastructure is operated solely for a specific organization, and is administrated by the organization or a third party. In a community cloud, the service is used by multiple organizations and made available only to those groups.**

Cloud Computing is considered to be a next step in the field of information technology. In this research work most of the researchers are trying to work on scheduling portion in order to gain maximum profit out of it. Therefore, in this work we have uses queue-based approach to identify and perform scheduling in an efficient manner.

Keywords— ***Cloud computing broker Scheduling, resource cost, resource management***

I. INTRODUCTION

Cloud computing has been described as synonym for distributed computing. It is a category computing which relies on sharing computing resources on pay per usage basis. Therefore, reducing cost of the infrastructure is of greatest importance. But reducing total charge of the cloud is not only the solution to give revenue but multiple features like response time and efficiency also plays a great role. Cloud computing in information communication technology is now coming to a place where a large horizon of industries stake holder coming to a single point of functions. Not only industry/organizations stakeholder but the users from different fields are using cloud computing at large scale. As we know cloud computing gives a shared pool of resources in order to collaborate with different users requests dynamic users' behaviour patterns also have a large impact on the efficiency of cloud computing. Therefore, in order to survive in the market each cloud vendor must learn their user behaviour and match their cycle. Cloud computing is a type of computing that relies on sharing computing resources rather than having local servers or personal devices to handle applications. The 'cloud' in cloud computing can be defined as the set of hardware, network, storage services and interface that combines to deliver aspects of computing as a service. Cloud service include the delivery of software, infrastructure &storage over the internet based on a user demand. In computer science Cloud computing is a synonym for distributed computing over a network and means the ability to run a program on many connected computers at the same time. The popularity of the term Cloud computing can be attributed to its use in marketing to sell hosted services in the sense of Application Service Provisioning that run Client server software on a remote location.

The three types most likely to be useful to consumers and businesses are Infrastructure as a Service, Platform as a Service, and Software as a Service.

1. Infrastructure as a Service (IaaS)

This is the foundation, or "base layer," of cloud computing, and it contains physical infrastructure such as servers, storage disks, and facilities. Organizations advantageous from pay-as-you-go, on-demand storage and web hosting, which can be easily scaled larger or smaller as need fluctuates. 2. Platform as a Service (PaaS)

This "middle layer" of cloud computing gives the operating system from which applications run. Here, the service operator gives a programming language and web server, which permit application developers [43] to develop and execute their software solutions.

In the middle, we have "Platform-as-a-Service," or "PaaS." At this service level, cloud providers deliver a computing platform typically including operating system, programming language execution environment, database, and web server.

3. Software as a Service (SaaS)

Finally, at the "top layer," we find software applications specifically developed for the internet. Here, consumers generally pay a monthly or yearly fee in order to use a certain software in the cloud (as opposed to traditional software, which requires a single, up-front cost for perpetual use). Because pricing is pay-per-user, organizations can quickly add or remove users without having to accordingly scale their associated platform [34] and infrastructure. This on-demand approach allows for rapid, efficient adjustments in staffing. Examples include Salesforce, Google Apps (Gmail, Google Calendar, Google Docs), and Microsoft Office 365.

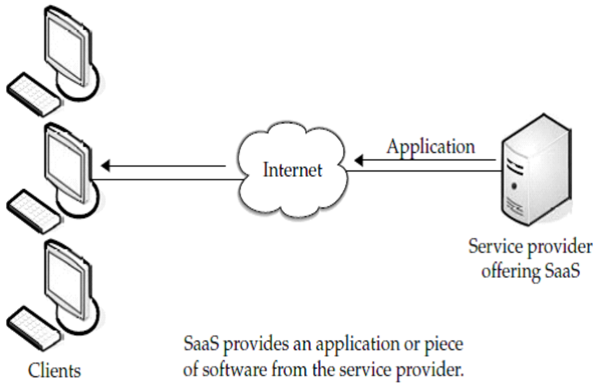
Application
Internet
Service provider offering SaaS
SaaS provides an application or piece of software from the service provider.
Clients

Fig (1) : Structure of Cloud

II. LITERATURE REVIEW

Huang and Subhlok[4] define some network transmission patterns, which include stable states, congestion states, and interrupt states. When a network transmission pattern resembles a denoted pattern, this information is utilized for pre-diction. This method has been compared against traditional methods like a simple moving average, exponential weighted moving average, and aggregate measured throughput. This approach performs as well or better than the other methodsin given scenarios Borzemski and Starczewski[5] focus on the regression based algorithms to predict time transfer. Similar to Huangand Subhloks work [4], this study develops a pattern state recognizer to perform the TCP throughput prediction of data transfers originated by clients.

There are some prior approaches aiming at building energy-efficient data centers, such as [6],[7], Bianchini, and Rajamony[8], and [9] ,[6] believe that Cloud Computing with Virtualization is a way to improve the energy efficiency of a datacenter.

[7] dynamically reconfigures a heterogeneous cluster to reduce energy consumption during off-peak hours. Bianchini, and Rajamony[8] identify the techniques for conserving energy in heterogeneous server clusters.

[9] show that using the dynamic voltage scaling (DVS) on each server node can achieve energy saving of 29%. Moreover, by turning off certain nodes based on workload achieves 40% energy saving. Petrucci et al. [10] propose a control mechanism for turning on/off server nodes according to the client connection number. This control mechanism maintains a pre-defined QoS (Quality of Service) while eliminating unnecessary power consumption.

[11] install several virtual machines (VMs) into a physical machine. In this approach, a network flow forecasting program acts as the manager for controlling the activation/deactivation of each node.[12] also proposes a system that uses network flow predictor to manage power consumption of server nodes while maintaining the required QoS. Besides designing a new energy-efficient hardware or data center, some researches mainly attack the energy efficiency issue regarding computation from the following perspectives: - Finds the most energy-efficient algorithm by comparing different algorithms, such as Bunse et al. [13]. Makes compilers to generate energy-efficient codes or use a energy-efficient library, such as R. Buyya, A. Beloglazov, J. Abawajy [14], E. Elnozahy, M. Kistler, R. Rajamony 15], and E. Pinheiro, R. Bianchini, E.V. Carrera [16], Bunse et al. [13] define a set of trend functions that chooses a sorting algorithm according to the given conditions. In their work, bubblesort, heapsort, insertionsort, mergesort, quicksort, selectionsort, shakersort, and shellsort are evaluated. Insertion sort is identified as the most energy-efficient sorting algorithm in this work, if the number of input items is large enough. [14] propose AcovSA (Analysis) of Compiler Options via Simulated Annealing) that can find a

good set of compiler options for a particular CPU and software.

E. Elnozahy, M. Kistler, R. Rajamony [15] tune the settings of register file with some code profiles. The main challenge of adopting such mechanism is the necessity of modifying ISA (Instruction Set Architecture). E. Pinheiro, R. Bianchini, E.V. Carrera [16] propose an energy feature library that is developed with many energy-saving techniques. Even recompilation sometimes causes unexpected interruptions.

III. ISSUSES IN CLOUD COMPUTING DURING SCHEDULING

Cloud computing is a vast emerging IT field that is in rapid form of expansion but cloud computing security risks, Privacy issues and other cloud computing threats are cited as the most substantial roadblock for cloud computing Uptake.

Security: Although the cloud computing service provider will provide with security essentials likes data storage and transmission encryption, authentication and authorizations but vulnerability of remote data access, Session ridding, virtual machine escape, Data storage, Criminal hackers, thieves and corrupt employees causes a lot of concern.

Reliability: Some people worry also about whether a cloud service provider is financially stable and whether their data storage system is trustworthy. Most cloud providers attempt to mollify this concern by using redundant storage techniques, but it is still possible that a service could crash or go out of business, leaving users with limited or no access to their data. A diversification of providers can help alleviate this concern, albeit at a higher cost.

Ownership: Once data has been relegated to the cloud, some people worry that they could lose some or all of their rights or be unable to protect the rights of their customers. Many cloud providers are addressing this issue with well-crafted user-sided agreements. That said, users would be wise to seek advice from their favorite legal representative. Never use a provider who, in their terms of service, lays any kind of ownership claim over your data.

Data Backup: Cloud providers employ redundant servers and routine data backup processes, but some people worry about being able to control their own backups. Many providers are now offering data dumps onto media or allowing users to back up data through regular downloads.

Data Portability and Conversion: Some people are concerned that, should they wish to switch providers, they may have difficulty transferring data. Porting and converting data are highly dependent on the nature of the cloud provider's data retrieval format, particular in cases where the format cannot be easily discovered. As service competition grows and open standards become established, the data portability issue will ease, and conversion processes will become available supporting the more popular cloud providers. Worst case, a cloud subscriber will have to pay for some custom data conversion.

IV. PROBLEM STATEMENT

Cloud computing that provides cheap and pay-as-you-go computing resources is rapidly gaining momentum as an alternative to traditional IT Infrastructure. As more and more users delegate their jobs to cloud providers, Service Level Agreements (SLA) between users and providers emerge as a key aspect. Due to the dynamic features of the cloud, continuous monitoring on Quality of Service (QoS) attributes is important to enforce SLAs. Also, various other features such as trust (on the cloud provider) come into consideration, particularly for enterprise customers that may outsource its critical data. This complicated nature of the cloud landscape warrants a sophisticated means of managing SLAs. This paper proposes a mechanism for managing SLAs in a cloud computing environment using the Web Service Level Agreement (WSLA) framework, developed for SLA monitoring and SLA enforcement in a Service Oriented Architecture (SOA).

We provide a brief summary of literature survey provides that with traditional task allocation systems in cloud the cost/profit ratio becoming very less. Since the fluctuating electricity cost is very large for distinct data centers present in various geographical positions. Therefore, an algorithmic efficiency must be taken into account with other parameters. To deal the situation of high cost due to SLA violations, geographical conditions etc. We present a pre-analyzer which pre-treats resources and requests before allocation.

Latency: Latency is defined as a delay flanked by transfers of data from one location to another. These delays annoying the consumers.

Throughput: It can be defined as a allocating of number of virtual machines per unit of time successfully.

Response time: It is a time when the task request submitted until the first response is produced.

Fairness: Without any unfairness, All the requests from tasks should be treated as same way for an allocation of a node. Therefore, performance is definitely one of the major concerns in using existing scheduling algorithm.

Performance factor (PF): The Performance is depending on an allocation of the resources. The correct resources must be allocating to a correct virtual machine. The major factor like performance, cost, time of execution depends upon allocation of a resources and this further depends upon scheduling of a resources.

V. PROPOSED WORK

In the proposed work, the model as a network bipartite graph $G = (L \cup V, E)$, where L denotes the set of data centers, V denotes the location of customers. For instance, V can be the set of access networks to which customers are connected. Denote by $E \subseteq L \times V$ the communication paths between customers and data centers. We also assign constant weights dlv to denote the network latency between a data center $l \in L$ and a client location $v \in V$. In our framework, we consider a discrete-time system model where time is divided into multiple time periods called reconfiguration periods corresponding to the timescale at which server placement and routing decisions are made. We assume that there is an interval of interest $K = \{0, 1, 2, ..., K\}$ that consists of K+1 periods. Let $N = \{1, 2, ..., n\}$ denote the set of SPs. We assume that at time $k \in K$, each customer location $v \in V$ has demand Dvk in terms of average arrival rate of requests from location v at time k. For simplicity, we assume that all the servers leased by each SP have identical size and functionality. For instance, a server can be a virtual machine (VM) that runs a specific application image. We define the state variable xl k $\in R+$ as the number of servers owned by the SP at location $l \in L$ at time k. To simplify the model, we assume that xl k can take continuous values rather than discrete values. This assumption is reasonable for large-scale services that require tens or hundreds of servers, where the weight of each individual server in the overall solution is small. In this case, we can always obtain a feasible solution by rounding up the continuous values to the nearest integer values. Based on this assumption, we can further decouple xl k by defining xlv k $\in R+$ as the number of servers at location l serving demand from $v \in V$.

VI. CONCLUSION AND FUTURE SCOPE

In recent years, some researchers have focused on the problem of resource management and performance control in data center [1, 2]. However, nowadays, most of those methods can not sufficiently adapt to complex cloud environment. These research efforts usually assume the system as the equilibrium state, and employ the method of average value analysis which is not sufficiently precise. E.g., aiming at increasingly complex computing system, in [1], IBM propose the conception of autonomic computing, whose definition is a technology which can realize self-management of system with the least manual intervention. Based on this conception, the authors in [2] propose that utility function is employed as objective function of self-management resources, which can be used to decide appropriate behaviour of every component according to utility value.

In [3], the authors propose a dynamic capacity allocation resolving model which is focused on multi-tier network applications, which can decide the number of resources should be allocated to every tier of application services, and judge the proper time at which these resources are allocated with the combination of prediction method. There are several works [3], [8], [14], [19] that use live migration as a resource provisioning mechanism, but all of them consider policy based heuristic algorithms to live migrate VMs, which is difficult in the presence of conflicting goals such as improving performance and reducing cost. For example, if the need for a policy change arises, the algorithm must be changed.

Migration of virtual machines is a well-organized system used to implement cost saving and load balancing in virtualized cloud computing data center. In this paper, we study the request allocation of multiple virtual machines from experimental perspective and investigate different resource reservation methods in the cost saving process as well as other complex migration strategies such as parallel migration and workload-aware migration.

REFERENCES

[1] Seth, B., Dalal, S., Jaglan, V., Le, D. N., Mohan, S., & Srivastava, G. (2020). Integrating encryption techniques for secure data storage in the cloud. Transactions on Emerging Telecommunications Technologies.

[2] Pamlin, D. (2008) The Potential Global CO2 Reductions from ICT Use: Identifying and Assessing the Opportunities to Reduce the First Billion Tonnes of CO2, Vol. May. WWF, Sweden.

[3] Zou and Liu Data Centre Energy Forecast Report. Final Report, Silicon Valley Leadership Group, July.

[4] Huang and Subhlok Metrics to Characterise Data Centre & IT Equipment Energy Use. Proc. Digital Power Forum, Richardson, TX, USA, September.

[5] Borzemski and Starczewski ORGs for scalable, robust, privacy-friendly client cloud computing. IEEE Internet Comput., September, 96–99.

[6] Fan, X., Weber, W.-D. and Barroso, L.A. (2007) Power provisioning for a warehousesized computer, Proc. 34th Annual Int. Symp. Computer Architecture, San Diego, CA, USA, June 9–13, 2007. pp. 13–23. ACM, NewYork.

[7] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. Soman,L. Youseff, and D. agorodnov, “The eucalyptus open-source cloud-computing system,” in Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid- Volume 00, pp. 124–131,2009.

[8] Bianchini, and Rajamony “Virtual infrastructure management in private and hybrid clouds,” IEEE Internet Computing, pp. 14–22, 2009.

[9] Jindal U., Dalal S., Dahiya N. A combine approach of preprocessing in integrated signature verification (ISV), International Journal of Engineering & Technology, Vol. 7, No. 1.2, 2018, pp. 155–159.

[10] K. Ye, X. Jiang, D. Ye, and D. Huang, “Two Optimization Mechanisms to Improve the Isolation Property of Server Consolidation in Virtualized Multi-core Server,” in Proceedings of 12th IEEE International Conference on High Performance Computing and Communications, pp. 281–288, 2010.

[11] S. Arora and S. Dalal, "Hybrid algorithm designed for handling remote integrity check mechanism over dynamic cloud environment", International Journal of Engineering & Technology, pp. 161-164, 2018

[12] Dalal S, Jaglan D V, Sharma D K K (2014). Designing architecture of demand forecasting tool using multi-agent system. International Journal of Advanced Research in Engineering and Applied Sciences, 3(1): 11–20

[13] R. Nathuji, K. Schwan, Virtualpower: coordinated power management in virtualized enterprise systems, ACM SIGOPS Operating Systems Review 41 (6),pp. 265–278,2007.

[14] R. Buyya, A. Beloglazov, J. Abawajy, Energy-efficient management of data center resources for cloud computing: a vision, architectural elements, and open challenges, in: Proceedings of the 2010 International Conference on Parallel and Distributed Processing Techniques and Applications, PDPTA 2010, Las Vegas, USA, 2010. 39

[15] E. Elnozahy, M. Kistler, R. Rajamony, Energy-efficient server clusters, Power-Aware Computer Systems, pp. 179–197,2003.

[16] E. Pinheiro, R. Bianchini, E.V. Carrera, T. Heath, Load balancing and unbalancing for power and performancee in cluster-based systems, in: Proceedings of the Workshop on Compilers and Operating Systems for Low Power, pp.182–195, 2001.

[17] S. Arora, S. Dalal and R. Kumar, "A Variant of Secret Sharing Protected with Poly-1305", Recent Advances in Computational Intelligence. Springer Cham, pp. 107-119, 2019.

[18] S. Yeo, H.-H. Lee, “Using Mathematical Modeling in Provisioning a Heterogeneous Cloud Computing Environment,” Computer, vol. 44, no. 8, pp. 55-62, 2011.

[19] S. Dalal and U. Jindal, "Performance of integrated signature verification approach: Review," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 3369-3373.

[20] L. Zhou, Y. Zhang, K. Song, W. Jing, and A. V. Vasilakos, “Distributed Media-Service Scheme for P2P-based Vehicular Networks,” IEEE Transactions on Vehicular Technology, vol. 60, no. 2, pp. 692-703, 2011.

[21] S. Arora and S. Dalal, "Trust Evaluation Factors in Cloud Computing with Open Stack", Journal of Computational and Theoretical Nanoscience, pp. 5073-5077, 2019.

[22] X. Song, B.-P. Paris, “Measuring the size of the Internet via importance sampling,” IEEE Journal on Selected Areas in Communications, vol. 21, no. 6, pp. 922-933, 2003.

[23] Malik Meenakshi, Nandal Rainu, Dalal Surjeet, Jalglan Vivek and Le DacNhuong 2021 Driving Pattern Profiling and Classification Using Deep Learning, Intelligent Automation & Soft Computing 28, 887-906.

[24] J.S. Chase, D.C. Anderson, P.N. Thakar, A.M. Vahdat, R.P. Doyle, Managing energy and server resources in hosting centers, in: Proceedings of the 18th ACM Symposium on Operating Systems Principles, ACM, New York, NY, USA, pp. 103–116, 2001.

[25] R. Raghavendra, P. Ranganathan, V. Talwar, Z. Wang, X. Zhu, No ‘‘power’’ struggles: coordinated multi-level power management for the data center, SIGARCH Computer Architecture News 36 (1) ,pp. 48–59,2008.

[26] Sharma, D., Sharma, K., & Dalal, S. (2014). Optimized load balancing in grid computing using tentative ant colony algorithm. International Journal of Recent Research Aspects, 1(1), 35–39.

[27] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat. Enforcing performance isolation across virtual machines in xen. In Proceedings of the ACM/USENIX International Conference on Middleware (Middleware), 2006.

[28] S. Arora and S. Dalal, "An Optimized Cloud Architecture for Integrity Verification", Journal of Computational and Theoretical N anoscience, pp. 5067-5072, 2019.

[29] D. Ongaro, A. L. Cox, and S. Rixner. Scheduling I/O in virtual machine monitors. In Proceedings of ACM International Conference on Virtual Execution Environments (VEE), 2008.

[30] Meenakshi Malik, Rainu Nandal, Surjeet Dalal, Vivek Jaglan and Dac-Nhuong Le, Deriving Driver Behavioral Pattern Analysis and Performance Using Neural Network Approaches, Intelligent Automation & Soft Computing, vol.32, no.1, 87-99, 2022, DOI:10.32604/iasc.2022.020249

[31] A. Verma, P. Ahuja, A. Neogi, pMapper: power and migration cost aware application placement in virtualized systems, in: Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware, Springer, pp. 243–264, 2008.