ISSN XXXX XXXX © 2016 IJESC

Research Article Volume 7 Issue No. 1

Survey on Cloud Infrastructure Resource Allocation for Big Data Applications

Sonali R. Mayne[1], Prof. S. D. Satav[2]
Department of Computer Engineering[1], Department of Information Technology[2]
JSPM's Jayawantrao Sawant College of Engineering, Hadapsar Pune, India

Abstract:
The System utilizes the technique BRA algorithm to combine different types of workloads. It is used to compute the unevenness in the utilization of multiple resources on a server. This algorithm consists of three parts such as load prediction, hot spot mitigation, and green computing. Cloud computing is sold on demand on the basis of time constrains basically specified in minutes or hours. Thus scheduling should be made in such a way that the resource should be utilized efficiently. In cloud platforms, resource allocation (or load balancing) takes place at two levels. First, when an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, attempting to balance the computational load of multiple applications across physical computers. Second, when an application receives multiple incoming requests, these requests should be each assigned to a specific application instance to balance the computational load across a set of instances of the same application.

Keywords: Cloud infrastructure, Big Data, Resource Allocation.

I. INTRODUCTION

With the development of Internet technologies and increasing demands of computer applications, Cloud Computing came as a multi-service provider that shares information, software, and open resources within the Internet-based environment. In October 2007, cloud computing was first introduced to the public through a cooperation between two computing companies, I.B.M and Google. After that, this new concept brought a variety of impacts and changes to numerous fields that were relevant to information technology (IT). Over the years, distributed environments have evolved from shared community platforms to utility-based models; the latest of these being Cloud computing. This technology enables the delivery of IT resources over the Internet [2], and follows a pay-as-you-go model where users are charged based on their consumption. There are various types of Cloud providers [2], each of which has different product offerings. They are classified into a hierarchy of as-a-service terms: Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). This paper focuses on IaaS Clouds which offer the user a virtual pool of unlimited, heterogeneous resources that can be accessed on demand. Moreover, they offer the flexibility of elastically acquiring or releasing resources with varying configurations to best suit the Requirements of an application. Even though this empowers the users and gives them more control over the resources, it also dictates the development of innovative scheduling techniques so that the distributed resources are efficiently utilized. Although this ecosystem has evolved around *public clouds* — commercial cloud providers that offer a publicly accessible remote interface for creating and managing VM instances within their proprietary infrastructure — interest is growing in open source cloud computing tools that let organizations build their own IaaS clouds using their internal infrastructures. These *private cloud* deployments' primary aim isn't to sell capacity over the Internet through publicly accessible interfaces but to give local users a flexible and agile private infrastructure to run service workloads within their administrative domains. Private clouds can also support a *hybrid cloud* model by supplementing local infrastructure with computing capacity from an external public cloud. Private and hybrid clouds aren't exclusive with being public clouds; a private/hybrid cloud can allow remote access to its resources over the Internet using remote interfaces, such as the Web services interfaces that Amazon EC2 uses. Here Virtual Machine (VM) performance is an additional challenge presented by Cloud platforms. VMs provided by current Cloud infrastructures do not exhibit a stable performance in terms of execution times. In fact, Schad report an overall CPU performance variability of 24% on Amazon's EC2 Cloud. The shared nature of the infrastructure as well as virtualization and the heterogeneity of the underlying non-virtualized hardware are some of the reasons behind such variability. This may have a significant impact when scheduling workflows on Clouds and may cause the application to miss its deadline. Many scheduling policies rely on the estimation of task runtimes on different VMs in order to make a mapping decision. This estimation is done based on the VMs computing capacity and if this capacity is always assumed to be optimal during the planning phase, the actual task execution will most probably take longer and the task will be delayed. This delay will also impact the task's children and the effect will continue to escalate until the workflow finishes executing.

II. LITURETURE SURVEY

A. *Real-Time Constrained Task Scheduling in 3D Chip Multiprocessor to Reduce Peak Temperature*

In this paper, we propose an online thermal prediction model for 3D chip. Using this model, we present a task scheduling algorithm based on rotation scheduling to reduce the peak temperature on chip. We consider the data Dependencies, especially the inter-iteration dependencies which are not well considered in most of the current thermal-aware task scheduling algorithms.

B . A Multimodal Approach for Evolutionary Multi-objective Optimization: MEMO

We have developed a constraint handling methodology that is well suited to the niching strategy to solve constrained multiband many-objective optimization problems. Results on two to 10-objective test problems and on several practical design problems from automotive and manufacturing industries have been compared with other state-of-the-art EMO methodologies and comparable results have been reported. The ability of a single-objective multimodal approach to find hundreds of Pareto-optimal solutions in a single simulation using an evolutionary algorithm reliably and consistently on many constrained and unconstrained problems remains as a hallmark achievement of this study.

C. EnReal: An Energy-Aware Resource Allocation Method for Scientific Workflow Executions in Cloud Environment.

Cloud platform expansion will make the energy consumption a big concern. In this paper, we propose an Energy-aware Resource Allocation method, named*EnReal*, to address the above challenge. Basically, we leverage the dynamic deployment of virtual machines for scientific workflow executions. Specifically, an energy consumption model is presented for applications deployed across cloud computing platforms, and a corresponding energy-aware resource allocation algorithmis proposed for virtual machine scheduling to accomplish scientific workflow executions. Experimental evaluation demonstrates that the proposed method is both effective and efficient.

III. EXISTING SYSTEM

The problem of mapping resources adaptively so that the resource demands of virtual machines are met in the cloud computing environment, while the number of physical machines used is minimized. So, Physical machine is overloaded and can lead to degraded performance of its virtual machines. On another hand, if the resource utilization of active server is too low, while the server is turned on resulting unnecessary use of power for big data applications. We tend to create the subsequent contributions. Overload avoidance: The capability of a PM ought to be ample to satisfy the resource wants of all VMs running thereon. Otherwise, the PM is full and might cause degraded performance of its VMs. inexperienced computing: the quantity of PMs used ought to be decreased as long as they'll still satisfy the requirements of all VMs. Idle PMs are often turned off to save lots of energy. we tend to develop a resource allocation system that may avoid overload within the system effectively whereas minimizing the quantity of servers used.

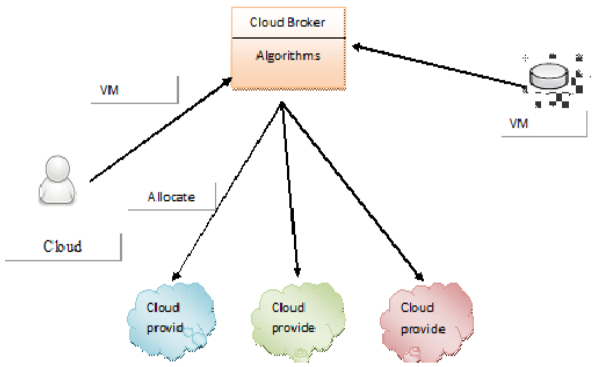

Figure. 1.Existing System

IV. PROPOSED SYSTEM

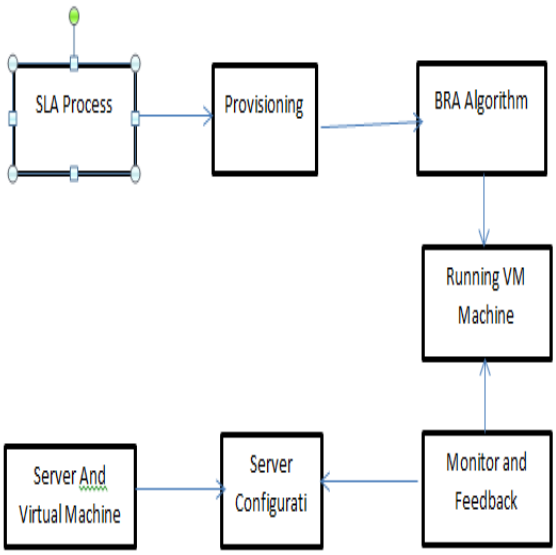

Figure. 2. Proposed System

The infrabody structure information parcelling problem in cloud s includes three main constrains, which are price, carrying into action, and availableness. The optimization root for information deployment is to achieve the highest functioning and handiness with the lowest cost. In stream cloud deployment solutions, various data culture medium s can be used for different aim. Different types of data medium have different costs, performance, and availableness respectively. In superior general, the higher the performance and availableness are, the higher the cost will be. In this paper, we use VMs, as the basic workings node in cloud infrastructure, to quantify data allocation problem. In general, supposing that all the VMs are with the same form , the more amount is, the higher performance and availability are, while the higher the cost will be. Meanwhile, the topological structure of data allocation has last relation with performance and availability. Because of the communication cost among working VMs, the farther the length between them is, the lower performance will be. However, the farther the aloofness is, the higher availability and security will be. In last, the cost, performance, and availability interplay with others in form of the topological structure. Heterogeneous data is an important Characteristic of cloud-based practical application. Consequently, the VMs used in cloud-based big data applications are not with the same configuration. They are with various operating organization, C.P.U., retentivity, networking bandwidth, and geographic location. As a result, we need to take heterogeneity into consideration, which data allocation problem extremely complicated to solve.

V. CONCLUSION

In this paper, we first analyzed the relations among the cost, performance, and availability of one cloud-based big data application, and built three models. Based on these three models we proposed BRA algorithm to obtain the optimal solution meeting all requirements. Then we designed and implemented a complete approach to allocate resources of big data application running on cloud. Finally, we perform three

sets of SLAs to verify the feasibility of our approach, and compared it with seven other approaches to show the effectiveness.

VI. REFERENCES

[1] http://www.gartner.com/it/page.jsp?id=1035013

[2] "Amazon elastic compute cloud (Amazon EC2), http://aws.amazon.com/ec2/."

[3] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho,R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proc. of the ACM Symposium on Operating Systems Principles (SOSP'03)*, Oct. 2003.

[4] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live migration of virtual machines," in Proc. of the Symposium on Networked Systems Design and Implementation(NSDI'05), May 2005

[5] X. Zhou, S. Gandhi, S. Suri, and H. Zheng, "ebay in the sky: strategy-proof wireless spectrum auctions," in Proc. of the 14th ACM Intl. Conf. on Mobile Comp. and Networking, 2008, pp. 2–13.

[6] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," technical report, Univ. of California, Berkeley, Feb.2009.

[7] Z. Kong, C.-Z. Xu, and M. Guo, "Mechanism design for stochastic virtual resource allocation in non-cooperative cloud systems," inProc. 4th IEEE Intl. Conf. on Cloud Computing, 2011, pp. 614–621

Author

Asst Prof. Sandip Satav received the M.E (CSE/IT) degree from Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, MAH, and India in 2004. He is currently working as Asst. Professor with Department of Information Technology, Jayawantrao Sawant College of Engineering, Pune, MAH, and India. His research interests include Image Processing, Networking.