ISSN:0975-9646

Improve Server Utilization and Achieving Green Computing in Cloud

M.Rajeswari[1], M.Savuri Raja[2], M.Suganthy[3]

[1 2] *Master of Technology,* [3]*Assistant Professor*
Department of Computer Science & Engineering,
Dr. S.J.S Paul Memorial College of Engineering and Technology, Puducherry – 605 502.

Abstract— **Cloud computing is a usage of very large scalable and virtualized resources in a dynamic way over the internet. Due to the rapid growth of cloud environment usage many tasks require to be executed by the available resources. At the same time it should be possible to achieve better performance, optimizing the servers, reduce migration, support green computing, better resource utilization etc. So resource allocation using virtual machine plays a most important role in cloud environment because it should allocate proper resources to various machines to get maximum benefit.**

Keywords—**Cloud computing, skewness, Resource Management, Virtualization, Green computing.**

I. INTRODUCTION

Cloud computing is a new technology currently being studied in the academic world [1]. The definition of the cloud computing from the Gartner: “A style of computing where massively scalable IT-related capabilities are provided as a service across the internet to multiple external customers using internet technologies. “Cloud computing is the delivery of computing services over the Internet .Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, web mail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

A. Service Levels Of Cloud Computing

Cloud service delivery is divided among three archetypal models and various derivative combinations. The three fundamental classifications are often referred to as the “SPI Model, “where ‘SPI’ refers to Software, Platform or Infrastructure (as a Service).

1) Cloud Software as a Service (SaaS): The capability provided to the consumer is to use the provider’s applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based email). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user specific application configuration settings.

2) Cloud Platform as a Service (PaaS): The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.

3) Cloud Infrastructure as a Service (IaaS): The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).

B. Deployment of Cloud Services

Regardless of the service model utilized (SaaS, PaaS, or IaaS) there are four deployment models for cloud services, with derivative variations that address specific requirements.

1) Public Cloud: A cloud is called a "public cloud" when the services are rendered over a network that is open for public use. Technically there may be little or no difference between public and private cloud architecture, however, security consideration may be substantially different for services (applications, storage, and other resources) that are made available by a service provider for a public audience and when communication is effected over a non-trusted network. Generally, public cloud service providers like Amazon AWS, Microsoft and Google own and operate the infrastructure and offer access only via Internet (direct connectivity is not offered).

2) *Private Cloud:* The cloud infrastructure is operated solely for a single organization. It may be managed by the organization or a third party, and may exist on-premises or off-premises. Undertaking a private cloud project requires a significant level and degree of engagement to virtualize the business environment, and requires the organization to reevaluate decisions about existing resources. When done right, it can improve business, but every step in the project raises security issues that must be addressed to prevent serious vulnerabilities.

Self-run data centres are generally capital intensive. They have a significant physical footprint, requiring allocations of space, hardware, and environmental controls. These assets have to be refreshed periodically, resulting in additional capital expenditures. They have attracted criticism because users "still have to buy, build, and manage them" and thus do not benefit from less hands-on management, essentially "[lacking] the economic model that makes cloud computing such an intriguing concept".

3) *Community Cloud:* The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, or compliance considerations). It may be managed by the organizations or a third party and may exist on-premises or off-premises.

4) *Hybrid Cloud:* The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds). It is important to note that there are derivative cloud deployment models emerging due to the maturation of market offerings and customer demand.

An example of such is virtual private clouds a way of utilizing public cloud infrastructure in a private or semi-private manner and interconnecting these resources to the internal resources of a consumers' datacenter, usually via virtual private network (VPN) connectivity. Varied use cases for hybrid cloud composition exist. For example, an organization may store sensitive client data in house on a private cloud application, but interconnect that application to a billing application provided on a public cloud as a software service. This example of hybrid cloud extends the capabilities of the enterprise to deliver a specific business service through the addition of externally available public cloud services.

By utilizing "hybrid cloud" architecture, companies and individuals are able to obtain degrees of fault tolerance combined with locally immediate usability without dependency on internet connectivity. Hybrid cloud architecture requires both on premises resources and off-site (remote) server-based cloud infrastructure.

II. RELATED WORK

In recent years, cloud computing has been emerging as the next big revolution in both computer networks and Web provisioning. Because of raised expectations, several vendors, such as Amazon and IBM, started designing, developing, and deploying cloud solutions to optimize the usage of their own data centers, and some open-source solutions are also underway, such as Eucalyptus and OpenStack.

Cloud architectures exploit virtualization techniques to provision multiple Virtual Machines (VMs) on the same physical host, so as to efficiently use available resources, for instance, to consolidate VMs in the minimal number of physical servers to reduce the runtime power consumption.VM consolidation has to carefully consider the aggregated resource consumption of co-located VMs, in order to avoid performance reductions and Service Level Agreement (SLA) [2] violations.

While various works have already treated the VM consolidation problem from a theoretical perspective, we focuses on it from a more practical viewpoint, with specific attention on the consolidation aspects related to power, CPU, and networking resource sharing. Moreover, proposes a cloud management platform to optimize VM consolidation along three main dimensions, namely power consumption, host resources, and networking.

System virtualization is becoming pervasive and it is enabling important new computing diagrams such as cloud computing. Live virtual machine (VM) migration is a unique capability of system virtualization which allows applications to be transparently moved across physical machines with a consistent state captured by their VMs. Although live VM migration is generally fast, it is a resource-intensive operation and can impact the application performance and resource usage of the migrating VM as well as other concurrent VMs. However, existing studies on live migration performance are often based on the assumption that there are sufficient resources on the source and destination hosts, which is often not the case for highly consolidated systems. As the scale of virtualized systems such as clouds continue to grow, the use of live migration becomes increasingly more important for managing performance and reliability in such systems.

Therefore, it is key to understand the performance of live VM migration under different levels of resource availability, addressing this need by creating performance models for live migration which can be used to predict a VM's migration time given its application's behavior and the resources available to the migration. A series of experiments were conducted on Xen to profile the time for migrating a DomU VM running different resource intensive applications while Dom0 is allocated different CPU shares for processing the migration. The results show that the VM's migration time is indeed substantially impacted by Dom0's CPU allocation whereas the coefficient of determination generally higher than 90%.

III. VIRTUALIZATION

Virtualization is a broad term that refers to computing elements running on a virtual basis rather than on a real basis, in order to simplify management, and optimize resource solutions. The users can use the same cost to build a more suitable space, and thus save costs, and maximize utilization of space to play. This concept of replanning to achieve maximum utilization of ideas with limited fixed resources according to different needs is called virtualization technology in the field of IT. Virtualization technology can expand the capacity of the hardware, and simplify software re-configuration process. The virtualization technology is a key technology for cloud computing. Cloud computing is a wide virtualization pool of resources; it allows users to easily use and access through the Internet; that is, such resources into the form of a "service", sent over the network to the user, so the users use the service according to personal needs.

On a cloud computing platform, dynamic resources can be effectively managed using virtualization technology. The subscribers with more demanding SLA can be guaranteed by accommodating all the required services within a virtual machine image and then mapping it on a physical server. This helps to solve problem of heterogeneity of resources and platform irrelevance.

Load balancing of the entire system can be handled dynamically by using virtualization technology where it becomes possible to remap virtual machines (VMs) and physical resources according to the change in load [3]. Due to these advantages, virtualization technology is being comprehensively implemented in cloud computing. However, in order to achieve the best performance, the virtual machines have to fully utilize its services and resources by adapting to the cloud computing environment dynamically. The load balancing and proper allocation of resources must be guaranteed in order to improve resource utility [4]. Thus, the important objectives of this to determine how to improve resource utility, how to schedule the resources and how to achieve effective load balance in a cloud computing environment.

IV. Skewness Algorithm

Skewness is a measure of the asymmetry or unevenness of the probability distribution. A distribution may either be positively or negatively skewed. We introduce the concept of skewness to compute the unevenness in the utilization of multiple resources on a server. Let n be the number of resources we consider and r be the utilization of the ith resource. We define the resource skewness of a server p as,

$$\text{Skewness} = \sqrt{\sum_{i=1}^{n}\left(\frac{r_i}{\bar{r}} - 1\right)^2}$$

Where,

‘$\bar{r}$’ is the average utilization of all resources for server p. In practice, not all types of resources are performance critical and hence we only need to consider bottleneck resources in the above calculation.

A. Analysis of Skewness Algorithm

The concept of skewness is to quantify the unevenness in the utilization of multiple resources on a server. By minimizing the skewness, we can improve the overall utilization of server resources. Let n and m be the number of PMs and VMs in the system, respectively. The number of resources (CPU, memory, I/O, etc.) that need to be considered is usually a small constant (e.g., 3 or 4). Thus the computation of the skewness and the temperature metrics for a single server takes a constant amount of time. During load prediction, we need to apply the FUSD algorithm to each VM and PM. The skewness algorithm consists of three parts such as load prediction, hot spot mitigation, and green computing.

1) Load Predication: The skewness algorithm executes repeatedly to evaluate the resource allocation status based on the predicted future resource demands of VMs. The utilization of any of its resources is above a hot threshold we define it as hot spot. This indicates that the server is overloaded and hence some VMs running on it should be migrated away. The temperature of a hot spot p as the square sum of its resource utilization beyond the hot threshold. The temperature of a hot spot reflects its degree of overload. If a server is not a hot spot, its temperature is zero.

The system defines a server as a cold spot if the utilizations of all its resources are below a cold threshold. This threshold indicates that the server is mostly idle and a system to turn off to save energy. By doing this, the actively used servers (i.e., APMs) in the system are below a green computing threshold. A server is actively used means it has at least one VM running. Otherwise, it is inactive. Finally, the proposed system define the warm threshold to be a level of resource utilization that is sufficiently high to justify having the server running but not so high as to risk becoming a hot spot in the face of temporary fluctuation of application resource demands. Different types of resources can have different thresholds.

2) Hot Spot Mitigation: In this the system sorts the list of hot spots in descending temperature. The goal of the system is to eliminate all hot spots if possible. Otherwise, keep their temperature as low as possible. For each server p, we first decide which of its VMs should be migrated away. We sort its list of VMs based on the resulting temperature of the server if that VM is migrated away.

We aim to migrate away the VM that can reduce the server’s temperature the most. In case of ties, the system selects the VM whose removal can reduce the skewness of the server the most.

2) Green Computing: When the resource utilization of all active servers is too low, some of them can be turned off to save energy. This is handled in our green computing algorithm. The challenge here is to reduce the number of active servers during low load without sacrificing performance either now or in the future. The algorithm is invoked when the average utilizations of all resources on active servers are below the green computing threshold.

We sort the cold spots in ascending order based on their memory size. Since the system needs to migrate away all its VMs before we can shut down an under-utilized server, we define the memory size of a cold spot as the aggregate memory size of all VMs running on it. Recall that our model assumes all VMs connect to share back-end storage. Hence, the cost of a VM live migration is determined mostly by its memory footprint.

B. Cloud Applications

Cloud computing is the delivery of computing as a service rather than a product, whereby shared resources, software and information are provided to users over the network. Cloud computing providers deliver application via the Internet, which are accessed from web browser, while the business software and data are stored on servers at a remote location.

Cloud computing can provide three kinds of service modes, including IaaS, PaaS and SaaS.

SaaS is a service provided to client in terms of applications running on the cloud computing infrastructure hosted by the service providers.

PaaS refers to services which provide high-level integrated environment to design, build, run, test, deploy and update the applications created by client using development language and tool say Java, python, .net etc. provided by the service providers to the cloud infrastructure.

IaaS refers to the services provided to the users is to lease the processing power, storage, network and other basic computing resources, with which users can deploy and run any software including operating systems and applications.

V. System Architecture

The architecture Fig.1 represents the design of the dynamic resource allocation for cloud computing environment, which consists of N servers each server consists of two virtual machines (VM) those are connected to the VM scheduler is connected to the internet to distribute the resources dynamically to the clients, the clients are accessing resources through the internet.

Virtual machine (VM) is a software implementation of computing environment in which operating system or program can be installed and run. Each VM encapsulate one or more applications such as Web server, remote desktop, DNS, Mail, Map/Reduce, etc. The memory usage within a VM, however, is not visible to the hypervisor. One approach is to infer memory shortage of a VM by observing its swap activities. Unfortunately, the guest OS is required to install a separate swap partition. Furthermore, it may be too late to adjust the memory allocation by the time swapping occurs.

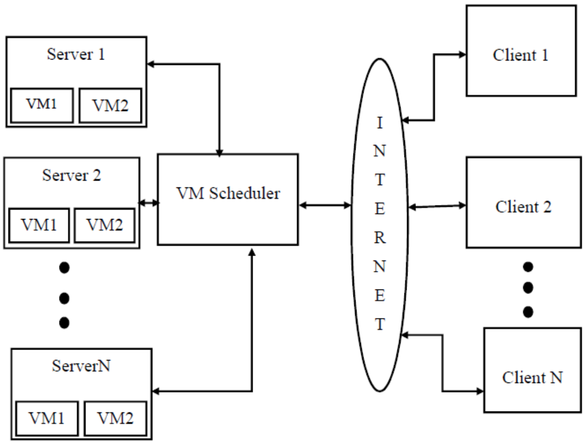

Fig .1 System Architecture

The statistics collected at each server are forwarded to the VM scheduler. The VM Scheduler is invoked periodically and receives the resource demand history of VMs, the capacity and the load history of PMs, and the current layout of VMs on servers. The scheduler has several components. The predictor predicts the future resource demands of VMs and the future load of servers based on past statistics. We compute the load of a server by aggregating the resource usage of its VMs.

VI. Conclusion

Dynamic resource allocation is growing need of cloud providers for more number of users and with the less response time. Cloud Computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the internet. We use the skewness metric to combine VMs with different resource characteristics appropriately so that the capacities of servers are well utilized. Our algorithm achieves both overload avoidance and green computing for systems with multi-resource constraints.

Acknowledgment

I wish to thank all the people who gave me a support from stage the idea was conceived. I would like to thank Ms.Suganthy, my project guide for their helpful comments and her suggestions. I express my thanks to friends who always stood as the helping and guiding support for me.

References

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

[2] N. Bobroff, A. Kochut, and K. Beaty, “Dynamic placement of virtual machines for managing sla violations,” in Proc. of the IFIP/IEEE International Symposium on Integrated Network Management (IM’07), 2007.

[3] P. Braham, 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] L. Cherkasova, D. Gupta, and A. Vahdat, “When virtual is harder than real: Resource allocation challenges in virtual machine based it environments,” Technical Report HPL 2007-25, February 2007.