ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering

Vol. 1, Issue 1, March 2013

A Study on Cloud Resource Management Techniques

D.Mahendran, M.Gopi, S. Priyadharshini, R.Karthick

Assistant Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore, India

ABSTRACT: Cloud computing allows cloud users to scale up and down their resource usage based on the requirements. Numerous cloud models come from resource multiplexing. Cloud computing relies mainly on sharing of the resources. As the need for resources increase there exists a need for the management of resources that are to be used. In this paper, we present a survey of resource management for cloud environment. The surveyportrays some of the resource management techniquesand the various metrics that are used to evaluate the allocation of resources.

KEYWORDS*:* Cloud computing, resource management, resource allocation, resource management schemes.

I. INTRODUCTION

Cloud computing is a way of enabling on-demand network access in order to share the resources such as storage, bandwidth, software, etc. Fig.1 shows basic cloud computing architecture. Cloud computing technology is a new concept of providing dramatically scalable and virtualized resources. It implies a service oriented architecture, reduced information technology overhead for the end-user, great flexibility, reduced total cost of ownership and many other things [1][3].

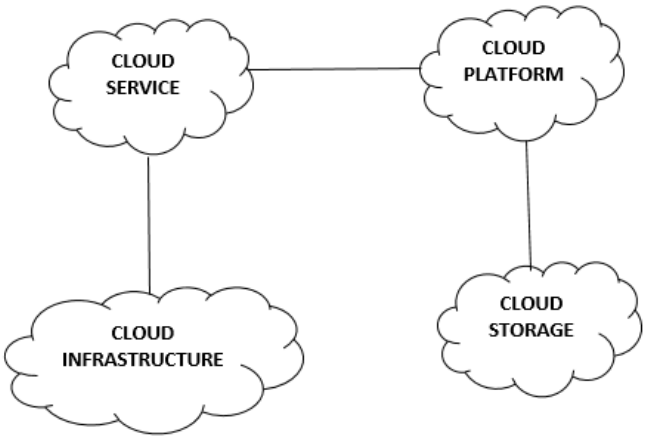

Fig. 1 Basic cloud architecture

A. Resource Management Strategies

In [3] Cloud computing that is based on resources acquired on demand is generating a good deal of interest among service providers and consumers. It is aiming to analyze the resource allocation and reallocation (load balancing) methods that are already present in the cloud environment and their bedrocks. Some issues of resource management are follows:

Identity management

Every enterprise will have its own identity management system to control access to information and computing resources. Cloud providers either integrate the customer's identity management system into their own infrastructure, using federation or SSO technology, or provide an identity management solution of their own.

Physical and personnel security

Suppliers make certain that physical machines are sufficiently secure and that access to these machines and also all related client data is not only controlled but that access is recognized.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering

Vol. 1, Issue 1, March 2013

Availability
Cloud suppliers guarantee clients that they will have usual and expected access to their data and applications.

Application security
Cloud providers ensure that applications available as a service via the cloud are secure by implementing testing and acceptance procedures for outsourced or packaged application code.

Privacy
Providers ensure that all critical data (credit card numbers) is masked and that only authorized users have access to data in its entirety. Digital identities and credentials must be protected as should any data that the provider collects or produces about customer activity in the cloud.

II. SURVEY OF RESOURCE MANAGEMENT

A. SLA Based Resource Allocation In Autonomic Environments

The service providers and their customers negotiate utility based Service Level Agreement (SLA) to determine the costs and penalties on the base of the achieved performance level [4]. The dispatcher can also decide to turn ON or OFF servers depending on the system load using resource allocation scheduler for such multi-tier autonomic environments so as to maximize the profits associated with multiple class SLAs. Autonomic systems maintain and adjust their operations in the face of changing components and the goal is to meet the application requirements while adapting IT architecture to workload variations. Network dispatcher manages autonomic components anddynamically determines the best use of resources on the base of a short-term load prediction A resource allocator is used for autonomic multi-tier environments. The main components of the network dispatcher [10, 11] are monitor, a predictor and a resource allocator. The system monitor measures the workload and performance metrics of each application, identifies requests from different customers and estimates requests service times. The predictor forecasts future system load conditions from load history and the allocator determines the best system configuration and applications to server's assignment. The design of a resource allocator is to maximize the revenue while balancing the cost of using the resources. The resource allocator can establish: (i) The set of servers to be turned ON depending on the system load, (ii) the application tiers to server assignment, (iii) The request volumes at various servers and (iv) The scheduling policy at each server. We model the problem as a mixed integer nonlinear programming problem and develop heuristic solutions based on a local search approach. The neighborhood exploration is based on affixed-point iteration (FPI) technique, which iteratively solves scheduling and a load balancing problem by implementing gradient method.

B. Resource Allocation Algorithms for Virtualized Service Hosting Platforms

Commodity clusters are used routinely for deploying service hosting platforms. Due to hardware and operation costs, clusters need to be shared among multiple services. Crucial for enabling such shared hosting platforms is virtual machine (VM) technology. The system proposes a formulation of the resource allocation problem in shared hosting platforms for static workloads with servers that provide multiple types of resources. This dynamic allocation of computing capacity is enabled by the virtualization of resources. This formulation makes it possible to compute abound on the optimal resource allocation. We propose several classes of resource allocation algorithms, which we evaluate in simulation. We are able to identify an algorithm that achieves average performance close to the optimal across many experimental scenarios. A formulation of the problem that supports a mix of QoS and best-effort scenarios, and that attempts to maximize a generic objective function, the minimum yield. We have proposed and evaluated several classes of algorithms over a wide range of simulation scenarios. We have found that performing a binary search over the yield and solving the resource allocation problem for a fixed yield using a vector packing algorithm is the best approach. Vector packing algorithms that reason on the sum of the resource needs of the services are the most effective. Among these algorithms the Chose Pack vector packing algorithm from [5] runs quickly and is the most effective.

C. Fast Transparent Migration For Virtual Machines

The design and implementation of a system uses virtual machine technology to provide fast, transparent application migration, Migrate unmodified applications on unmodified mainstream Intel x86-based operating system, including Microsoft Windows, Linux, Novell NetWare and others[7]. Neither the application nor any clients communicating with the application can tell that the application has been migrated. Fast transparent migration can improve global system

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering

Vol. 1, Issue 1, March 2013

utilization by load balancing across physical machines, and can improve system serviceability and availability by moving applications off machines that need servicing or upgrades.

Virtual Machine Migration

The migration must be apparent to the guest operating system, remote clients of the virtual machine applications running on the operating system. It should appear to all parties involved that the virtual machine did not change its location. The only perceived change should be a brief slowdown during the migration and a possible improvement in performance after the migration because the VM was moved to a machine with more available resources. The vmkernel schedules the VMM for each virtual machine and allocates and manages the resources needed by the virtual machines.

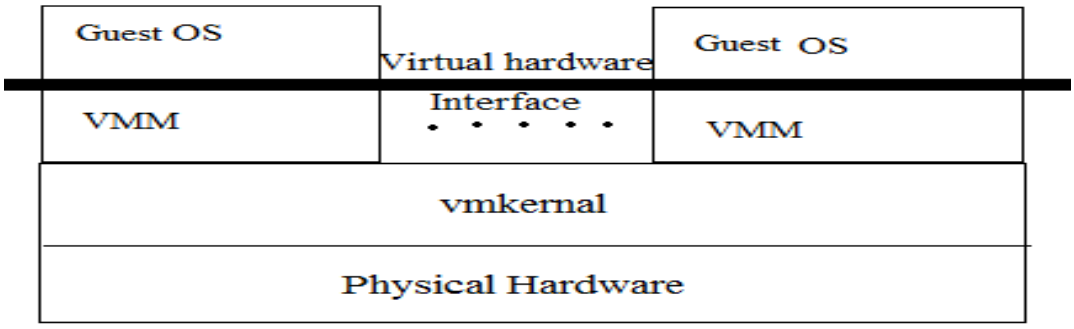

Fig. 2 VM platform layers in VMware ESX Server

A. PowerNap: Eliminating Server Idle Power

PowerNap is an energy conservation approach for heavy load environment. Our goal is to obtain the minimizing idle power and low transiton time. PowerNap eliminate the idle power waste in enterprise blade server. PowerNap operate on low efficiency region (see fig. 6), it uses the RAILS provides high conservation energy. RAILS mean Redundant Array for Inexpensive load sharing. PowerNap with RAILS will reduce average server power consumption by 74%. PowerNap eliminates idle power in server by quickly transitioning in and out of ultra-low power state. It improves power conversion efficiency and reduces the cost [8].

B. VirtualPower: Coordinated Power Management

It expresses how to integrate power management mechanism and policies with the virtualized technology. The goal is to support the isolated and independent operation and to control and globally coordinate effects of diverse power management.
An implementation of Virtual Power Management (VPM) for the Xen hypervisor addresses this challenge by provision of multiple system-level abstractions including VPM states, channels, mechanisms, and rules. The multi core platforms are high-light resulting improvements in online power management capabilities, including minimization of power consumption with little or no performance penalties and the ability to throttle power[9].Infrastructure is to support rich and effective policies for allocations across different vms. It prevents power viruses. It is used for Intel Core micro architecture.

III. METRICS ASSOCIATED WITH RESOURCE MANAGEMENT STRATEGIES

In [3] Resource management is achieved through some sort of load balancing among the participating nodes. There are some metrics that will help to evaluate the efficiency of each load balancing method. Load balancing techniques in cloud environment, consider various parameters like performance, scalability, response time, throughput, resource utilization, fault tolerance, migration time and associated overhead.

- **Overhead Associated:** determines the amount of overhead involved while implementing load balancing algorithms. It is composed of overhead due to movement of tasks, inter-process and inter-processor communication. This metric should be minimized so that a load balancing technique can work efficiently.
- **Throughput:** it is used to calculate the number of tasks whose execution has been completed. It should be high to improve the performance of the system.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering

Vol. 1, Issue 1, March 2013

- **Performance**:is used to check the efficiency of the system. It has to be improved at a reasonable cost e.g. reduce response time while keeping some acceptable delays.
- **Resource Utilization:**it is used to check the utilization of resources in a system. It should be optimized for an efficient load balancing.
- **Scalability:** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.
- **Response Time:**is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. It should be minimal to improve efficiency.
- **Fault Tolerance:** is the ability of an algorithm to perform uniform load balancing in spite of arbitrary node or link failure. Every system is expected to be highly fault tolerant.
- **Migration time :** is the time to migrate the jobs or resources from one node to other. It should be minimized in order to enhance the performance of the system.
- **Energy Consumption**: Load balancing helps in avoiding overheating by balancing the workload across all the nodes of a Cloud, hence reducing energy consumption.

IV. CONCLUSION

Nowadays cloud computing technology is increasingly being used in enterprises and business markets. In cloud environments, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. It tries to analyze the resource allocation strategies based on various matrices and it points out that some of the strategies are efficient than others in some aspects.

REFERENCES

[1] "Overview of Cloud Computing," http://en.wikipedia.org/ Feb, 2011.
[2] "Introduction of service model of cloud computing," http://en.wikipedia.org/ Feb, 2011.
[3] K. Rasmi and V. Vivek, "Resource Management Techniques in Cloud Environment - A Brief Survey," International Journal of Innovation and Applied Studies (IJIAS '13), April 2013.
[4] DaniloArdagna, Marco Trubian, Li Zhang, "SLA based resource allocation in autonomic environments", journal of parallel and distributed computing, 2007.
[5] Mark Stillwell, David Schanzenbach, Frédéric Vivien, Henri Casanova, "Resource allocation algorithms for virtualized service hosting platforms" journal of parallel and distributed computing, 2010.
[6] 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," Proc. ACM Symp. Operating Systems Principles(SOSP *'03),* Oct. 2003.
[7] M. Nelson, B.-H. Lim, and G. Hutchins, "Fast Transparent Migration for Virtual Machines," *Proc. USENIX* Ann. Technical Conf., 2005.
[8] D. Meisner, B.T. Gold, and T.F. Wenisch, "Powernap: Eliminating Server Idle Power," Proc. Int'l Conf. Architectural Support for Programming Languages and Operating Systems (ASPLOS '09), 2009.
[9] R. Nathuji and K. Schwan, Virtualpower: Coordinated Power Management in Virtualized Enterprise Systems," Proc. ACMSIGOPS Symp. Operating Systems Principles (SOSP '07), 2007.