International Journal of Computer Applications Technology and Research
Volume 2– Issue 6, 731 - 737, 2013

Survey on Dynamic Resource Allocation Strategy in Cloud Computing Environment

N.Krishnaveni
Dept. of CSE
Erode Sengunthar Engineering College
Thudupathi, India

G.Sivakumar
Dept. of CSE
Erode Sengunthar Engineering College
Thudupathi, India

Abstract-Cloud computing becomes quite popular among cloud users by offering a variety of resources. This is an on demand service because it offers dynamic flexible resource allocation and guaranteed services in pay as-you-use manner to public. In this paper, we present the several dynamic resource allocation techniques and its performance. This paper provides detailed description of the dynamic resource allocation technique in cloud for cloud users and comparative study provides the clear detail about the different techniques.

Keywords: Cloud computing, Dynamic Resource Allocation, Cloud users, Resources, Data center, Virtual machine

1. INTRODUCTION

Cloud computing allows customers to scale up and down their resources based on needs. Cloud computing technology makes the resources as a single point of access to the client and cost is pay per usage. Cloud computing is a computing technology, where a pool of resources are connected in private and public networks and to provide these dynamically scalable infrastructure for application. Cloud computing is not application oriented and this is a service oriented. It offers the virtualized resources to the cloud users. Cloud computing provide dynamic provisioning and thus can allocate machines to store data and add or remove the machines according to the workload demands. Cloud computing platforms such as, those provided by Microsoft, Amazon, Google, IBM. Cloud computing is an environment for sharing resources without the knowledge of the infrastructure and can makes it possible to access the applications and its associated data from anywhere at any time.

Cloud environment provide the four types of cloud.

- Public cloud
- Private cloud
- Hybrid cloud
- Community cloud

Cloud computing offers three types of services

- Software as a service(Saas)
- Platform as a service(Paas)
- Infrastructure as a service(IaaS)

Virtualization technology

Cloud computing is based on the virtualization technology. Virtualization technology is used to allocate the data center resources dynamically based on the application demands.

Virtualization having two types,

- Para virtualization
- Full virtualization

Live migration

Virtual machine live migration technology makes it possible to mapping between the virtual machines (VMs) and the physical machines (PMs) while applications are running. Live migration increase the resource utilization and provide the better performance result.

2. RESOURCE ALLOCATION

In cloud computing, Resource allocation is the process of assigning available resources to the needed cloud applications. Cloud resources can be provisioned on demand in a fine-grained, multiplexed manner. In cloud the resource allocation is based on the infrastructure as a service (IaaS).In cloud platforms, resource allocation takes place at two levels:

- when an application is uploaded to the cloud, the load balancer assigns the requested instances to physical computers, to balance the computational load of multiple applications across physical computers
- When an application receives multiple incoming requests, these requests should be assigned to a specific application instance to balance the

computational load across a set of instances of the same application

Resource allocation techniques should satisfy the following criteria:

- Resource contention arises when two applications try to access the same resource at the same time
- Resource fragmentation arises when the resources are isolated. There would be enough resources but cannot allocate it to the needed application due to fragmentation.
- Scarcity of resources arises when there are limited resources and the demand for resources is high
- The multiple applications needed different types of resources such as cpu,memory,I/O devices and the technique should satisfy that request
- Over provisioning of resources arises when the application gets surplus resources than the demanded one

3. RESEARCH ISSUES IN DYNAMIC RESOURCE ALLOCATION TECHNIQUES

In this paper we have analyzed some of the dynamic resource allocation techniques in cloud environment

Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds

In Existing system generate the more messages for a single request. The proposed system using the SOC and it achieves the maximized resource utilization and it also delivers optimal execution efficacy.

SOC:

SOC connect a large number of desktop computers on the internet by P2P network. Each participating computer act as a resource provider and resource consumer.

SOC having two main issues:

- Locating a qualified node to satisfy a user task's resource demand with bounded delay
- To optimize a task's execution time by determining the optimal shares of the multi-attribute resources to allocate to the tasks with various QoS constraints, such as the expected execution time

Algorithm:

- Dynamic optimal proportional share
- Multi range query protocol

DOPS:

This algorithm used to redistribute available resources among running tasks dynamically, such that these tasks could use up the maximum capacity of each resource in a node, while each task's execution time can be further minimized.

Procedures:

Slice handler: It is activated to equally scale the amount of resources allocated to tasks.

Event handler: It is used for resource redistribution upon the events of task arrival and completion.

Multi Range Query Protocol:

This algorithm used to locate qualified nodes in the SOC environment; we design a fully-decentralized range query protocol, namely pointer-gossiping CAN (PG-CAN), DOPS to find the qualified resources with minimized contention among requesters based on task's demand. It is unique in that for each task, there is only one query message propagated in the network during the entire discovery.

Range query protocol proactively diffuses resource indexes over the network and randomly route query messages among nodes to locate qualified ones that satisfy tasks' minimal demands. To avoid possibly uneven load distribution and abrupt resource over-utilization caused by un-coordinated node selection process from autonomous Participants.

Dynamic Resource Allocation for Spot Markets in Clouds

As a demand of each VM type can fluctuate independently at run time, it becomes a problem to dynamically allocate data center resources to each spot market to maximize cloud provider's total revenue.

We present a solution to this problem that consists of 2 parts:

- Market analysis for forecasting the demand for each spot market
- A dynamic scheduling and consolidation mechanism that allocate resource to each spot market to maximize total revenue.

Cloud providers specify a fixed price for each type of VM offerings.

- When total demand is much lower than data center capacity, the data center becomes under-utilized, i.e., the cloud provider is to encourage customers to submit more requests.
- When total demand rises over the data center capacity, it is desirable for the cloud provider to motivate the customers to reduce their demand.

A promising solution to this problem is to use market economy to reshape the demand by dynamically adjusting the price of each VM type.

- When total demand is high, the mechanism raises the price to ensure resources are allocated to users who value them the most.
- When total demand is low, the mechanism lowers the prices and provides incentive for customers to increase their demand.

Dynamic resource allocation framework consists of the following components:

Market Analyzer:

- Analyze the market situation and forecast the future demand and supply level
- Predict the future demands use AR(auto regressive model)

Capacity planner:

Capacity planner decide the expected price of each VM

Different pricing schemes:
In the fixed pricing scheme, price of a VM type does not vary with the current supply and demand.
In the uniform pricing scheme, the price of a VM type is adjustable at run-time.
VM scheduler:

- Make online scheduling decision for revenue maximization
- Dynamic resource allocation policy has the multiple machine configuration and this will amplified when the demand pattern changes over the time.

Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment

Cloud computing allows business customers to scale up and down their resource usage based on needs. In this paper, using virtualization technology to allocate data center resources dynamically based on application demands and support green computing by optimizing the number of servers in use
Technique:

1. Virtualization technology

2. Skewness

Goals:

Overload avoidance: The capacity of a PM should be sufficient to satisfy the resource needs of all VMs running on it.

Green computing: The number of PMs used should be minimized as long as they can still satisfy the needs of all VMs. Idle PMs can be turned off to save energy.

Virtualization technology: This technology used to allocate datacenter resources based on the application demands.

Skewness: This is used to measure the unevenness multidimensional resource utilization of a server. To minimizing skewness, we can combine different types of workloads.
Skewness can be measured based on,
Hot spot: If the utilization of any of its resources is above a hot threshold. This indicates that the server is overloaded and hence some VMs running on it should be migrated away.
Cold spot: If the utilizations of all its resources are below a cold threshold. This indicates that the server is mostly idle and a potential candidate to turn off to save energy.
Achieve the goals to make the following contributions,
1. Develop a resource allocation system that can avoid overload in the system
2. skewness to measure the uneven utilization of a server.
3. Design a load prediction algorithm that can capture the future resource usages of applications accurately without looking inside the VMs.
Load prediction algorithm:
Exponentially weighted moving average (EWMA): Predict the CPU load and we measure the load every minute and predict the load in the next minute.

Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud

In cloud computing environment data analytics are the key applications and it should be account for the heterogeneity of environments and workloads. In cloud computing environment does not provide the fairness among jobs when multiple jobs share the cluster. In Proposed, Resource allocation on a data analytics system in the cloud to hold the heterogeneity of the underlying platforms and workloads and the architecture shows how to allocate resources to a data analytics cluster in the cloud.

Technique:

Data analytic cluster: Data analytics workloads have heterogeneous resource demands because some workloads may be CPU whereas others are I/O-intensive. In cloud various resource demands of data analytics workloads, we scale the cluster according to demands.

For scaling process the resource allocation strategy having two levels

(1) Divides machines into two pools - core nodes and accelerator nodes

(2) Dynamically adjusts the size of each pool to reduce cost or improve utilization.

Data analytic cloud contains the components are,
Core nodes: Host the data and computations
Accelerator nodes: This is added to the cluster temporarily when additional computing power needed.
Analytic engine:Runs the application both the pools
Cloud driver:

- This manages the nodes allocated to the analytic cloud and decides when to add/remove what type of nodes to/from which pool.
- The user submits the job to the cloud driver with the hints about the job.Cloud driver keeps the history of routinely processed query, this is used to estimate the submission query and update the hints provided.
- This is also monitor the storage system to estimate the incoming data rate. This will predict the resource requirements to process queries and to store data.
- Many productions query are submitted with tight deadlines, the cloud driver will add the nodes to the accelerator pool temporarily to handle the job rather than allocating more core nodes.

- When adding nodes, the cloud driver makes the decision on which resource container to use.

Priority Based Resource Allocation Model for Cloud Computing

Cloud computing is a model which enables on demand network access to a shared pool computing resources. A cloud environment consists of multiple customers requesting for resources in a dynamic environment with possible constraints. In existing system cloud computing, allocating the resource efficiently is a challenging job. The cloud does not show the Qos, SLA.

This paper proposed allocates resource with minimum wastage and provides maximum profit. The developed resource allocation algorithm is based on different parameters like time, cost, No of processor request etc.
Algorithm:

Priority algorithm:
Priority algorithm that mainly decides priority among different user request based on many parameters like cost of resource, time needed to access, task type, number of processors needed to run the job or task.

Resource Allocation Model:
In this model client send the request to the cloud server. The cloud service provider runs the task submitted by the client. The cloud admin decides the priority among the different users request.
Each request consists of different task and it have the different parameters such as ,
Time- computation time needed to complete the particular task,
Processor request- refers to number of processors needed to run the task. More the number of processor, faster will be the completion of task.
Importance- refers to how important the user to a cloud administrator (admin) that is whether the user is old customer to cloud or new customer.
Price- refers to cost charged by cloud admin to cloud users.

Survey on Resource Allocation Strategies in Cloud Computing

In cloud computing, the important problem is to manage the Qos and to maintain the SLA for cloud users that share cloud resources. In this paper proposed Dynamic resource allocation can be based on,
1. Topology Aware Resource Allocation (TARA)
2. Linear scheduling strategy
3.Dynamic resource allocation for parallel data processing

Topology Aware Resource Allocation (TARA):
This allocation scheme based on the Iaas cloud systems. IaaS systems are usually unaware of the hosted application's requirements and therefore allocate resources independently of its needs.
Overcome this problem this method using the prediction engine and genetic algorithm based search.
Prediction engine with lightweight simulator to estimate the performance of the given resource allocation.
Genetic algorithm based search technique that allows TARA to guide the prediction engine

Linear scheduling strategy:
Linear Scheduling performs tasks and resources scheduling respectively. Here, a server node is used to establish the IaaS cloud environment and KVM/Xen virtualization with LSTR scheduling to allocate resources which maximize the system throughput and resource utilization.

Dynamic resource allocation for parallel data processing:
This technique used to allocate and deallocate the resources from a cloud during job execution. Some of the particular task can be assigned to different types of virtual machines and these task are automatically instantiated and terminated during job execution.

Survey on Resource Allocation Strategies in Cloud Computing (2012)

In cloud computing, the important problem is to manage the Qos and to maintain the SLA for cloud users that share cloud resources.
In cloud computing there are many RAS techniques,

Execution time: Estimating the execution time for a job is a hard task for a user and errors are made very often. This paper proposed technique is matchmaking strategy and it is based on Any-Schedulability criteria for assigning jobs to resources in heterogeneous environment.

Policy: This is based on the two levels: security and processor.

Security policy proposed a decentralized user and virtualized resource management for IaaS by adding a new layer called domain in between the user and the virtualized resources. Based on role based access control (RBAC), virtualized resources are allocated to users through domain layer.

Processor policy for resource allocation means the job is allocates to the cluter,then the number of processor needed for the subsequent job allocation. The number of processors in each cluster is binary compatible. Job migration is required when load sharing activities occur.

Virtual machine:The dynamic availability of infrastructure resources and dynamic application demand, a virtual computation environment is able to automatically relocate itself across the infrastructure and scale its resources in cloud environment. This is also based on the load, cost, speed and the type of application.

Utility function: Dynamically manage VMs in IaaS by optimizing some objective function such as minimizing cost function, cost performance function and meeting QoS objectives. The objective function is defined as Utility

property which is selected based on measures of response time, number of QoS, targets met and profit.

Hardware Resource Dependency:Improve the hardware utilization, we propose the Multiple Job Optimization (MJO) scheduler. Jobs can be classified by hardware-resource dependency such as CPU, Network I/O, Disk I/O and memory bound. MJO scheduler can detect the type of jobs and parallel jobs of different categories. Based on the categories, resources are allocated.

Auction:In this method the cloud provider collects the user's proposals and determines the price. This not provides the more profit. This achieved by using market based resource allocation strategy. Dynamically adjust the resources in single VM according to various resource requirements of workloads.

SLA:RAS to focusing on SLA has driven user based QoS parameters to maximize the profit for SaaS providers. The mappings of customer requests in to infrastructure level parameters and policies that minimize the cost by optimizing the resource allocation within a VM.

A Dynamic Resource Allocation Methods for Parallel Data Processing

Nephele is the first data processing framework to explicitly exploit the dynamic and probably heterogeneous. In existing system the resource overload is high. The proposed system increases the efficacy of the scheduling algorithm for the real time cloud computing services. The algorithm utilizes the turnaround time utility efficiently by differentiating it into a gain function for a single task.

The algorithm assigns high priority for early completion task and less priority for abortions/deadlines.
Cloud computing performance can be improved by,

Associate each task with the time utility function (TUF).this is not important to measure the profit when completing a job in time but also account the penalty when a job is aborted or discarded.

In nephele architecture, the client submits the task job manager. Job manager allocate and deallocate VMs. VMS can be differentiated based on the instance type. For example,"m1.small" is a instance type means, it refers 1 cpu core, 1GB RAM, 128GB disk. The task manager receive tasks from the job manager at a time and decides how many and what type of instances job should be executed. The algorithm proves,

- Preemptive scheduling provides the maximum profit than the non-preemptive scheduling.
- Non-Preemptive scheduling provides the maximum penalty than the preemptive scheduling.

Efficient Idle Desktop Consolidation with Partial VM Migration

Idle desktop systems are frequently left powered, often because of applications that maintain network presence. Idle PC consumes up to 60% of its peak power desktop VM often large requiring gigabytes of memory. These VM creates bulk transfer and utilize server memory inefficiently. In existing technique using the ballooning method, this not ensures the quick resume and provides the strain to the network.

Proposed system: Using the partial VM migration technique. This migrates only the working set of an idle VM.it allows user applications to maintain the network presence while the desktop sleeps and to transfer the execution of an Idle VM and it fetches the VM's memory and disk state on –demand.

Partial migration leverages two insights:

- First, the working set of an idle VM is small, often more than an order of magnitude smaller than the total memory allocated to the VM.
- Second, rather than waiting until all state has been transferred to the server before going to sleep for long durations, the desktop can save energy by micro sleeping early and often, whenever the remote partial VM has no outstanding on-demand request for state.

Working set migration: when consolidating a VM from the desktop to the server, partial VM migration transfers memory state only as the VM requires for its execution.

State Access Traces

Mean Memory working set was only 165.63 MiB with standard deviation of 91.38MiB.

The mean size of disk access during idle times was 1.16 MiB with standard deviation of 5.75MiB.

Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective

Virtualization and VM migration capabilities enable the data center to consolidate their computing services and use minimal number of physical servers.
In previous works, the issue of SLA violation has not received thorough analysis. In this work, we devise an algorithm that will keep the migration time minimum as well as minimizing the number of migrations. This will play a major role in avoiding the performance degradation encountered by a migrating VM.

The main aim is to placing the virtual machine using the technique bin packing algorithm and gradient search technique. The heuristic based VM migration scenario is partitioned as follows:

- Determining when a physical server is considered to be overloaded requiring live migration of one or

more VMs from the physical server under consideration.

- Determining when a physical server is considered as being under loaded hence it becomes a good candidate for hosting VMs that are being migrated from overloaded physical servers.
- Selection of VMs that should be migrated from an overloaded physical server. VM selection policy (algorithm) has to be applied to carry out the selection process.
- Finding a new placement of the VMs selected for migration from the overload and physical servers and finding the best physical

This empirical study seeks to achieve the following goals:

- Carrying out the live migration of VMs in a manner that preserves free resources in order to prevent SLA violations
- Optimal utilization of resources
- Performing minimal number of migrations to the extent possible
- Efficient server consolidation through VM migrations

4. COMPARISON OF THE DYNAMIC RESOURCE ALLOCATION TECHNIQUE

TITLE	ADVANTAGE	PARAMETER RESULT
Dynamic Optimization of Multi-Attribute Resource Allocation in Self Organizing Cloud	Locating qualify nodes and optimize task execution time	Throughput Ratio: 60% improvement
Priority Based Resource Allocation Model for Cloud Computing	Resource wastage is minimized	Parameters: No.of users, Time to run, No.of processor, job type, User type
Dynamic Resource Allocation Using Virtual Machine for Cloud Computing Environment	Server overload is minimized	Migration of VM for resource Requirement
Survey on Resource Allocation Strategies in Cloud Computing(2013)	It should maintain the SLA and also manage the Qos	Strategies: Virtual machine,SLA,utility
Heterogeneity Aware Resource Allocation In Cloud	Provide the fairness among jobs when multiple jobs are submitted	The result is based on the Instance Type Ex:m1.small
Dynamic Resource Allocation for Parallel Data Processing in cloud	Overload is avoided	Gain utility: Preemptive>Non-Preemptive Penalty: Non Preemptive> preemptive
Efficient Idle Desktop Consolidation with Partial VM Migration	This migrates only working set of an idle VM	This can deliver the 85%to 104% of the energy saving compare full VM migration
Survey on Resource Allocation in Cloud Computing	This avoids the resource contention and scarcity of resources	Technique: Topology aware resource allocation
Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective	Less SLA violation and less performance degradation	Average: SLA violation Is reduced to 17.64 to 16.44
Dynamic Resource Allocation for Spot Market in Cloud	Total revenue is maximized	Income:15173.28 Loss:1083.63 NetIncome:14089.65

5. COMPARISON OF PERFORMANCE EVALUATION

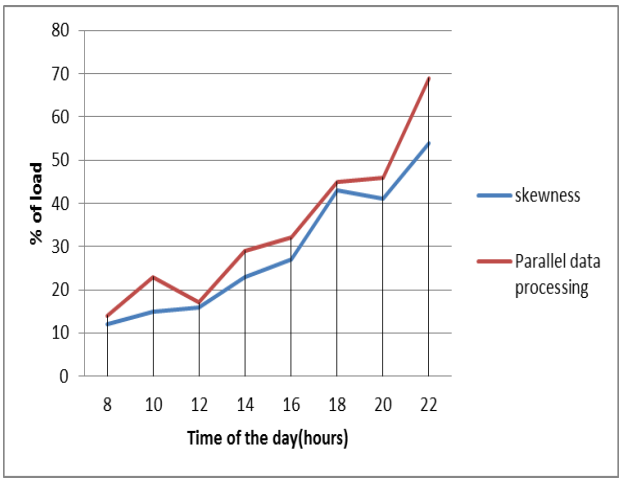

Fig.1 Overload of the server

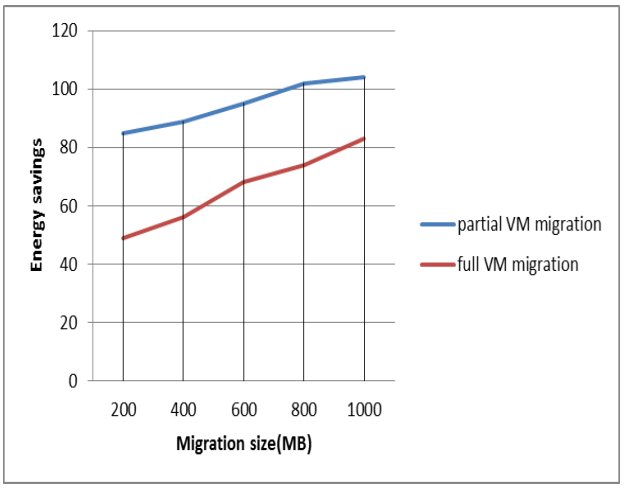

Fig.2 Energy saving based on VM migration

6. CONCLUSION

This paper addresses the theoretic study of various dynamic resource allocation techniques in cloud environment. The detail description of the techniques is summarized and also summarizes the advantages with parameters of the various techniques in cloud computing environment.

7. REFERENCES

[1] Ronak Patel, Sanjay Patel" Survey on Resource Allocation Strategies in Cloud Computing" International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 2, Feb- 2013

[2] K C Gouda, Radhika T V, Akshatha M," Priority based resource allocation model for Cloud computing" International Journal of Science, Engineering and Technology Research (IJSETR)Volume 2, Issue 1, January 2013.

[3] Gunho Leey, Byung-Gon Chunz, Randy H. Katzy," Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud", IJERT-2012.

[4] Nilton Bilay, Eyal de Laray, Kaustubh Joshi_, H. Andr´es Lagar-Cavilla ,Matti Hiltunen_ and Mahadev Satyanarayananz," Efficient Idle Desktop Consolidation with Partial VM Migration", Journal of computer application-2012.

[5] Venkatesa Kumar, V. And S. Palaniswami," A Dynamic Resource Allocation Method for Parallel data processing in Cloud Computing", Journal of Computer Science 8 (5): 780-788, 2012.

[6] Sheng Di and Cho-Li Wang," Dynamic Optimization of Multi-Attribute Resource Allocation in Self-Organizing Clouds", IEEE Transactions on parallel and distributed systems, - 2013.

[7] V.Vinothina, Dr.R.Sridaran, Dr.padmavathiganapathi," A Survey on Resource Allocation Strategies in Cloud Computing "International Journal of Advanced Computer Science and Applications, Vol. 3, No.6, 2012.

[8] Qi Zhang, Eren G¨urses, Raouf Boutaba, Jin Xiao," Dynamic Resource Allocation for Spot Markets in Clouds", Journal of computer science-2012.

[9] Zhen Xiao, Senior Member, IEEE, Weijia Song, and Qi Chen," Dynamic Resource Allocation Using Virtual Machines for Cloud Computing Environment", IEEE Transactions on parallel and distributed systems, vol. 24, no. 6, June 2013.

[10] Ts`epomofolo, R Suchithra," Heuristic Based Resource Allocation Using Virtual Machine Migration: A Cloud Computing Perspective", International Refereed Journal of Engineering and Science (IRJES) Volume 2, Issue 5(May 2013)

.