Volume 3, Issue 12, December 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Survey on Resource Management Strategies for Desktop-as-a-Service in Cloud Environment

Gandhi Kishan Bipinchandra*,
M.Tech(CE) Research Scholar,
RK University, India

Dr.Ajay Shanker Singh ,
Dept. of Computer Engg,
RK University, India

Abstract: - Cloud computing is an on demand technology as it offers dynamic and supple resource allocation for trustworthy and guaranteed services in pay as-you-use way to public. The specialty of this technology is that the any number of cloud services can be concurrently accessed by any number of users. In desktop-as-a-service cloud computing, user applications are executed in Virtual Desktop on remote servers. This offers great advantages in terms of usability and resource utilization; however, handling a large amount of clients in the most efficient manner poses important challenges. Especially deciding how many clients to handle on one server, and where to execute the user applications at each time is important. Assigning too many users to one server leads to customer dissatisfaction, while assigning too little leads to higher investments costs. We study different aspects to optimize the resource usage and customer satisfaction. Here in this paper a survey is carried out on the area of resource allocation strategies which tries to preserves the customer satisfaction to its maximum. The merits and demerits of each technique are also discussed.

Keywords: - cloud computing, Desktop-as-a-Service, service level agreements, resource overbooking, resource management strategies.

I. INTRODUCTION

The Cloud computing services such as Amazon's Elastic Compute Cloud, are widely available today, offering computing resources on demand. Thanks to such advances and ubiquitous network availability, the thin client computing paradigm is enjoying increasing popularity. Originally intended for wired LAN environments this paradigm is repeating its success in a mobile context. A study from ABI Research forecasts a US$20 billion turnover surrounding services directly associated with mobile cloud computing by the end of 2014. Clearly, when applications are offloaded, the mobile terminal only needs to present audiovisual output to users and convey user input to remote servers, considerably reducing the client device's computational complexity. Consequently, applications can run as-is, without requiring (many) scaled-down versions for mobile devices. Several popular applications, such as Google Docs and Microsoft Live, already execute on servers in the cloud. The ability to access applications in the cloud is referred to as software as a service (SaaS).There has been a rapid adoption of "cloud" platforms for online applications such as email, photo/video galleries and file storage in academia and industry. The next frontier for these user communities will be to transition their "traditional distributed desktops" that have dedicated hardware and software installations into "virtual desktop clouds" (VDCs) that are accessible via thin-clients. Moreover, in the not so distant future, we can envisage home users signing-up for virtual desktops (VDs) with a VD Cloud Service Provider (CSP) providing Desktop-as-a-Service (DaaS) as a utility.

Concept of desktop-as-a-service cloud computing is very attractive for several reasons, for example, lower client hardware investments are required, the end-user is no longer bothered with regular updates and often difficult installation and configuration of applications or anti-virus software, lower IT management costs for companies thanks to central management of desktops and applications. Furthermore, since the processing power of servers in the network is used, desktop-as-a-service cloud computing enables access to advanced applications (e.g., computer-aided design (CAD) applications) from any device, for example from a tablet PC. Mobile users would no longer need to use restricted mobile versions of their applications. Deliver desktop-as-a-services and Windows applications hosted on hypervisors such as Microsoft Hyper-V, Citrix Xen, VMware vSphere and more, to any remote mobile or desktop device and grant access by username, group, IP or MAC address

II. APPLICATION OF DESKTOP-AS-A-SERVICE:-

Current desktop-as-a-service computing deployments are typically operational in corporate local area network (LAN) environments, which are highly controlled environments offering fixed and stable bandwidth availability to a relative small, well-known user base. Extending desktop-as-a-service computing to wide area network (WAN) environments, which comprise a large, geographically distributed customer base, where users are potentially connected through unreliable wireless network connections, involves a number of novel challenges. Strategies are needed to improve resource utilization and/or customer satisfaction in WAN environments in the most efficient manner. Cloud computing [1] is an enabler for this kind of service. Unlike most current cloud services, the applications are not accessed through a web browser (e.g. Google Apps Cloud Service[2]), but through a thin client protocol (e.g., Microsoft Remote Desktop Protocol (RDP) [3] or Virtual Network Computing (VNC) [4]). This way, legacy applications must not be rebuilt to be offered by the envisioned service.

A. Latest RDP/RemoteFX Multimedia Support:-

Provide a rich remote desktop and application experience by supporting Microsoft's high definition RemoteFX accelerated multimedia protocol .

B. Enterprise Scalability:-

Allow more concurrent users to work within the same farm with Sites Management while enabling multiple administrators with different roles to simultaneously interact within the system through Role Based Administration.

C. Desktop-as-a-service Infrastructure (DI) Management for Major Hypervisors:-

Manage and administer core DI capabilities such as templates, pooling and more for multiple hypervisors

III. RESOURCE MANAGEMENTAND ITS NECCESITY

In desktop-as-a-service cloud computing, user applications are executed in Virtual Desktop on remote servers. This offers great advantages in terms of usability and resource utilization; however, handling a large amount of clients in the most efficient manner poses important challenges. Especially deciding how many clients to handle on one server, and where to execute the user applications at each time is important. Assigning too many users to one server leads to customer dissatisfaction, while assigning too little leads to higher investments costs. In cloud computing the resource management has got a cogent role in the performance of the whole process and the level of customer satisfaction provided by the process. But while providing the maximum customer satisfaction the service provider ought to be definite the profits that incur to them also. So the resource management ought to be efficient on both perspectives i.e. on the end user and the service provider point of view.

IV. RESOURCE MANAGEMENT STRATEGIES

Cloud computing that is rooted in resources acquired on demand is generating a lot of interest among service providers and consumers. Here in this section we are going to analyze the resource allocation and reallocation (load balancing) methods that are previously present in the cloud environment and their basic principle.

A. Autonomic Workload Provisioning (AWP)

Executing applications in the underlying resources has got two key steps in the case of cloud environment. The first step is called the VM provisioning. It consists of creating VM instances with the intent of hosting each application request and then the harmonizing of the specific features and the needs of the request. And the second step in this process is mapping and forecasting these requests onto distributed physical resources. That is known as resource provisioning. Most of the virtualized data centers currently provide a set of general purpose VM classes and they are provided with generic resource configurations. The specialty is that it quickly becomes insufficient to support the highly varied workloads.

In [5] it proposes a new autonomic workload provisioning that addresses the challenges of enterprise grids and clouds. The main aim of this mechanism is that to improve the
Resource utilization and which is achieved with the help of reducing the over provisioning. It can be reached through two levels.

In order to reduce the over provisioning caused by the difference between the virtual resources allocated to VM instances and those requested by the individual jobs a new mechanism is introduced. And this technique is base on decentralized online clustering, and it helps to characterize the resource requirement classes and it is used for proactive VM provisioning.

This paper also introduced another way of resolving the over provisioning problem which may be happened due to inaccuracies in client resource requests. This paper also explored the use of workload modeling techniques and their application. The mechanism for dynamic and
Decentralized VM provisioning monitors the flow of arriving jobs from different queues in a decentralized manner during ongoing analysis windows of duration in the order of the startup time of new VMs.

The decentralized clustering mechanism possesses several advantages. It has the capability of analyzing jobs across a dynamic set of distributed queues. It has no prior knowledge
Regarding the existence of the number of clustering classes and also it has the adaptively changing behavior towards the dynamic workloads and resources. The main drawback of this method is that it is not suitable for real-time application. It is because of the fact that the resource provisioning is done after the queuing of the requests. That is each request has to wait in the multiple queues before they get serviced. This causes some reduction in the QoS factor and the user satisfaction.

B. Linear Scheduling Strategy

The resource allocation is taken into consideration commonly the parameters like CPU utilization, memory utilization and throughput etc. The cloud environment has to take into reflection all these things for each of its clients and could provide maximum service to all of its clients. In [6] it suggests that when we are taking the scheduling of resources and tasks separately it imposes large waiting time and response time.

In order to overcome this shortcoming this paper introduces a new approach namely Linear Scheduling for Tasks and Resources (LSTR). Here scheduling algorithms primarily focus on the distribution of the resources along with the requestors which will make best use of the selected QoS parameters. The QoS parameter selected in this approach is the cost function.

The scheduling algorithm is designed based on the tasks and the available virtual machines together and specified as LSTR scheduling strategy. This is designed so as to maximize the resource utilization. Scheduling algorithm is carried out based on the prediction that the initial response to the request is made only after collecting the resource for a finite amount of time but not allocating the resource as they arrive. But dynamic allocation could be carried out by the scheduler dynamically on request for extra resources. This is obtained by the continuous evaluation of the threshold value within the system. Here the authors state that this approach is appropriate when we are considering the "shortest job first" rather than the FCFS approach. It is for the reason that the algorithm sorts the requests by excluding the arrival times. It only considers the "threshold" of the request for the scheduling purpose.

This approach has the advantage that it has an improved throughput and response time. But it is not appropriate for the interactive real-time areas because there is no consideration for the arrival time. For interactive applications the requests are treated in a "first come first serve" basis.

C. Pre-Copy Approach

Clark et al. [7], talks about the live migration of the virtual machines. In this paper they suggest that migration of the operating system instances across distinct physical hosts is a useful tool for the administrator of data centers and clusters. It also provides a separation between hardware and software and facilitates load balancing, fault management and low level system maintenance.

In "Pre-Copy Approach" pages of memory are iteratively copied from the source machine to the destination host and in addition there is a fact that all these things are done without ever stopping the execution of the system. Page level protection hardware is used to make sure that a consistent snapshot is transferred. For controlling the traffic of other running services a rate-adaptive algorithm is used. And during the final phase it pauses the virtual machine and copies any remaining pages to the destination and after that resumes the execution
there.

Franco et al. [8] put forward some of the drawbacks that are encountered in the above Mentioned approach . It points out that the conventional approach in [7] is inadequate because of the high RTTs and potential store and forward handling of virtual machines. It also points out that it will result in long forwarding chains. This will create a delay to the user experiences with the system.

D. MIPS Based Scheduling

One of the important requirements for a cloud computing environment is to provide reliable QoS. It can be able to define in terms of Service Level Agreements, it describes about such characteristics like maximal response time, minimal throughput, or latency delivered by the system.

In [9] it talks about a system comprises of a large-scale Cloud data center comprising of diverse physical nodes. Here each node has a CPU, which can be a multi core and the specialty is that its performance is defined in Millions Instructions Per Second (MIPS). The specialty of the MIPS is that it depends deeply on the instructions to be executed [10]. The usual way to measure CPU potential is FLOPS (Floating Point Operations per Second). And in FLOPS it will specify which type of instruction you are dealing with. In short "MIPS" cannot be a well-founded way to judge processors for their performance.

E. Match Making and Scheduling

Shikharesh et al. [12], suggests that the "Match making" is the first step and "scheduling" is second in the resource allocation in cloud environment. Matchmaking is the procedure for allocating jobs associated with user requests to resources selected from the available resource pool. Scheduling refers to determining the order in which jobs mapped to a specific resource are to be executed [12]. It also points out that there are some uncertainties that are associated with such type of "match making" and scheduling. They can be like

(1) Error Associated with Estimation of Job Execution Times

It is considered that estimating the execution time for a line of work is a very hard labor and faults may happen very often.

(2) Lack of Knowledge of Local Resource Management Policies

Matchmaking is challenging in cloud systems because the scheduling policy used at Every resource may not be known to the resource broker. Resource broker performs admission control for advanced reservations during the request to resource mapping. This negative condition happens because of the fact that the exact system configuration for a cloud may not be fully known during the time of system design or deployment. After all it may change many times during the lifetime of the entire system. So the method given in [12] is vulnerable to some sort of uncertainties that are explained above.

F. Just-In-Time Resource Allocation

In [14] it talks about the cost based workload provisioning and "just- in- time Resource allocation".

(1) Workload Prediction

Here the prediction of the workload on the application and estimation of the system behavior over the prediction horizon is using a performance model. Here optimization of the system behavior is carried on by taking into consideration the minimization of the cost incurred to the application.

(2) Just-in-time Resource Allocation

In this just in time resource allocation the three components of the cost function refer individually to the penalty for violation of SLA bounds, cost of leasing a machine, and cost of reconfiguring the application when machines are either leased or released. But for the look-ahead implementation of the time interval for each task need the implementation of recursive data structures. And the prediction of this look-ahead time also results in some prediction error [14].

V. COMPARISION

Here in this section it carries out a brief comparison between the resource management strategies discussed above. The merits and demerits of each method is mentioned. Table gives the overall summary of the comparisons made

Table I
COMPARISION BETWEEN DIFFERENT RESOURCE MANAGEMENT STRATERGIES

AUTHOR	**METHOD**	**MERITS**	**DEMERITS**
Quiroz et al. [5]	Autonomic Workload Provisioning	Maximum resource utilization, reduced over provisioning.	Queuing of requests, not suitable for real-time applications.
Abirami S.P.,Shalini Ramanathan [6]	Linear Scheduling Strategy	Improved throughput and response time.	Not suitable for interactive real time applications
Clark et al. [7]	Precopy Approach	Page level protection hardware	Long forwarding chains, delayed user experiences
Anton Beloglazov ,Rajkumar Buyya [9]	MIPS based reallocation	Depends on instructions to be executed.	Not the proper way for measuring the CPU performance.
Shikharesh Mujumdar [12]	Match making and scheduling	Cost effective, less delay	Uncertainties that are Associated with such type of "match making". Error Associated with Estimation of Job Execution Times. Lack of Knowledge of Local Resource Management Policies
Roy et al. [14]	Just-in-time Resource allocation	Cost effective	Prediction error and use of recursive data structures

VI. RESOURCE OVERBOOKING

Resource overbooking is nothing but reserving resources in advance. In [13] it gives a detailed description of the overbooking technique and what are the advantages that the customer can benefit from this technique in a cloud. It is more useful in the concept of virtualization, clearly say virtual desktops.

Resource overbooking is the technique that can establish an increase of the average utilization of hosts in a data center by reserving fewer resources than required in worst case. Since more virtual desktops can be allocated to a host, the cost for the service provider related to investment in hardware equipment, server maintenance cost and energy cost can be reduced. The risk parameter that is limiting the degree of overbooking is the risk to affect the user satisfaction [15]. Fiedler in this proposed method suggests a careful overbooking for network virtualization and it also obeys service level agreements (SLA) for full and limited availability. Full availability means the availability of all the required resources. Limited availability stands for the availability of certain share of required resources that are statically guaranteed at given degrees.

Urgaonkar et al. in [16] mentioned about how to maximize the revenue through overbooking. They suggest that provisioning cluster resources based on the worst case needs of the application results in low average utilization. it is because of the fact that average resource requirements of an application are normally smaller than its worst case requirements. And also the resources tend to idle when at times when the application does not utilize its peak reserved share. In [12] it summarizes that in shared hosting platforms techniques to overbook (i.e. under provision) resources in a guarded manner will outcome in revenue maximization through optimized usage.

VII. CONCLUSIONS

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. This paper summarizes different resource management strategies and its impacts in cloud system. Here also mentioned the concept of overbooking and its advantages and certain limiting factors. And it is found that the overbooking concept has got a close relationship with effective resource management in cloud environment.

REFERENCES

[1] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging it platforms: Vision, hype, and reality for de- livering computing as the 5th utility. Future Generation Computer Systems, 25(6):599-616, 2009.

[2] Google. Google apps cloud service. http://www.google.com/apps.

[3] Microsoft Corporation. Windows Remote Desktop Protocol (RDP).http://www.microsoft.com/ntserver/ProductInfo/terminal/tsarchitecture.asp.

[4] T. Richardson, Q. Sta_ord-Fraser, K.R. Wood, and A. Hopper. Virtual network com-puting. Internet Computing, IEEE, 2(1):33-38, 1998

[5] Quiroz A, Kim H, Parashar M, Gnanasam bandam N, SharmaN; “Towards workload provisioning for enterprise grids and clouds”. 10th IEEE/ACM international conference on grid computing, 2009. pp 50-57.

[6] Abirami S.P. ,Shalini Ramanathan; “Linear Scheduling Strategy for Resource allocation in Cloud Environment”;International Journal on Cloud Computing and Architecture,vol.2, No.1, February 2012.

[7] Christopher Clark, Keir Fraser, Steven Hand, Jacob GormHanseny, Eric July, Christian Limpach, Ian Pratt, AndrewWarfield; “Live Migration of Virtual Machines”, 2ndSymposium on Networked Systems Design and Implementation (NSDI) , May 2005.

[8] Franco Travostino, Paul Daspit, Leon Gommans, Chetan Jog,Cees de Laat, Joe Mambretti, Inder Monga,Bas vanOudenaarde, Satish Raghunath, Phil Wang; “Seamless Live Migration of Virtual Machines over the MAN/WAN” ;Elsevier Future Generation Computer Systems 2006.

[9] Anton Beloglazov ,Rajkumar Buyya ; “Energy Efficient Resource Management in Virtualized Cloud Data Centers”;10th IEEE/ACM International Conference on Cluster, Cloudand Grid Computing 2010.

[10] http://www.tomshardware.com/forum/285409-28-mips

[11] Stephen S. Yau , Ho G. An. “Adaptive Resource Allocation for Service-Based Systems”, International Journal of Software andInformatics ISSN 1673-7288, Vol.3, No.4, December 2009, pp.483–499.

[12] Shikharesh Mujumdar; “Resource management on cloud:Handling uncertainities in parameters and policies" CSI communications , 2011 , edn. pp.16-19.

[13] Lien Deboosere , Bert Vankeirsbilck ,Pieter Simoens , Filip DeTurck , Bart Dhoedt and Piet Demeester, “Efficient resource management for virtual desktop cloud computing”, Springer2012.

[14] Nilabja Roy, Abhishek Dubey and Aniruddha Gokhale ;“Efficient Autoscaling in the Cloud using Predictive Modelsfor Workload Forecasting”.

[15] Markus Fiedler, “On Resource Sharing and Careful Overbooking for Network Virtualization”, 20th ITC Special Seminar, May 2009.

[16] Bhuvan Urgaonkar, Prashant Shenoy and Timothy Roscoe, “Resource Overbooking and Application Profiling in Shared Hosting Platforms”, ACM Trans Internet Tecnology2009