ISSN: 2321-7782 (Online)

Volume 2, Issue 1, January 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Paper

Available online at: www.ijarcsms.com

Dynamic Resource Allocation Techniques in Cloud Computing

Pratik P. Pandya[1]
M.tech in Computer Engineering
School of Engineering
RK University
Rajkot - India

Hitesh A. Bheda[2]
M.E in Computer Engineering
School of Engineering
RK University
Rajkot - India

Abstract: Dynamic resource allocation is very much popular research area in cloud environment due to its live application in data center.Becasue of dynamic and heterogeneous nature of cloud, allocation of virtual machine is affected by various parameters like QOS, time consumption, cost, carbon effect etc. Grouping of virtual machine which is communicate to each other to execute one large request is comes into affinity group. Here we will study details of allocation method, affinity of virtual machine and how it will give good performance over non-affinity group and give some idea about new technique to improve performance which we will implement in future.

Keywords: Affinity Grouping; Resource Allocation; VM Packing; Virtual Cluster; Cloud Computing.

I. Introduction

The emergence of cloud computing (e.g., Amazon EC2) has led to a growing interest in deploying a wide variety of applications on shared computing environments. In particular, because of the relative abundance of resources and low cost of resource outsourcing, clouds are highly attractive for compute-intensive applications. The success of clouds has been driven in part by the use of virtualization as their underlying technology. Virtual machines (VMs) provide flexibility and mobility through easy migration, which enables dynamic mapping of VMs to available resources.

Virtual machines also provide performance isolation and security that facilitates multiplexing and utilization of shared resources. For these reasons, virtualization has also become popular in other domains such as scientific and high-performance computing. Virtualization has become a crucial technology in cloud computing, in which applications, such as parallel computing applications and multi-tier e-business web applications, are encapsu-lated within multiple virtual machines (VMs), and dynamically assigned to a pool of physical machines (PMs) for provisioning cloud services. The execution of application jobs inside VMs generates a large amount of communications or data exchanges across these VMs. With the increasingly growing demands of handling cloud service provision tasks, the network, as a key infrastructure in a cloud datacenter, is sustaining a tremendous pressure. Due to the poor efficiency of network virtualization and resource allocation, the network bandwidth becomes a bottleneck in many existing virtualized datacenters, leading to the intensification of network congestion and performance degradation for communication or data intensive applications. In my review, I address a resource allocation problem in which VMs have resource demands and dependency across VMs is identified as affinity relationship. The rest of the paper is organized as follows: Section II contain basic information about resource allocation, related work discussed in section III, in section IV includes problem definition and in section V contain proposed work.

II. Resource Allocation

Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application. It requires the type and amount of resources needed by each application in order to complete a user job. The order and time of allocation of resources are also an input for an optimal RAS. An optimal RAS should avoid the following criteria as follows:

a) Resource contention situation arises when two applications try to access the same resource at the same time.

b) Scarcity of resources arises when there are limited resources.

c) Resource fragmentation situation arises when the resources are isolated.

d) Over-provisioning of resources arises when the application gets surplus resources than the demanded one.

e) Under-provisioning of resources occurs when the application is assigned with fewer numbers of resources than the demand.

From the perspective of a cloud provider, predicting the dynamic nature of users, user demands, and application demands are impractical. For the cloud users, the job should be completed on time with minimal cost. Hence due to limited resources, resource heterogeneity, locality restrictions, environmental necessities and dynamic nature of resource demand, we need an efficient resource allocation system that suits cloud environments. Cloud resources consist of physical and virtual resources. The physical resources are shared across multiple compute requests through virtualization and provisioning. The request for virtualized resources is described through a set of parameters detailing the processing, memory and disk needs. Provisioning satisfies the request by mapping virtualized resources to physical ones. The hardware and software resources are allocated to the cloud applications on-demand basis. For scalable computing, Virtual Machines are rented. The complexity of finding an optimum resource allocation is exponential in huge systems like big clusters, data centres or Grids. Since resource demand and supply can be dynamic and uncertain.

III. Related Work

There are various work done in Dynamic resource allocation using various techniques, in this section we will see one by one technique. Davide Tammaro said that computing job requests that are characterized by their arrival and teardown times, as well as a predictive profile of their computing requirements during their activity period. Assuming a prior knowledge of the predicted computing resources required by end-users, they propose and investigate several algorithms with different optimization criteria. However, prediction errors may occur resulting in some cases in the drop of one or several computing requests [1].Where Chandrashekhar S. Pawar mention that services execute by priority based where application is pre-emptable. An algorithm is divided into four steps where first work is distributed among working vm by load balancer then forming a task list based on priorities. On the third steps Cloud min-min scheduling (CMMS) used for scheduling and Priority Based Scheduling Algorithm (PBSA) used at the last [2]. Jason Sonnek noted minimization of communication overhead in virtualized platform using affinity-migration concept. In which they used bartering agent who can handle all things that done in virtual machine when any job is executed [3]. Yi Wei [4] said that as business processes and scientific jobs become more intricate, users may have unbalanced and evolving requirements that are associated to the specific underlying sub-job or service. They investigate the problem of managing virtual resources for service workflows in a cloud environment and propose an agent-based framework that makes resource management decisions that are customized to optimize different levels of the application. To facilitate the management process, they introduce and evaluate an adaptive workflow configuration algorithm to assign virtual machines to service (composed within a workflow) while aggregating virtual machines on appropriately-resourced physical machines. Venkatesa Kumar [5] proposed both algorithm of pre-emptive and non-preamble job then compare it with result as well a used time utility function (TUF) with nephel’s architecture.

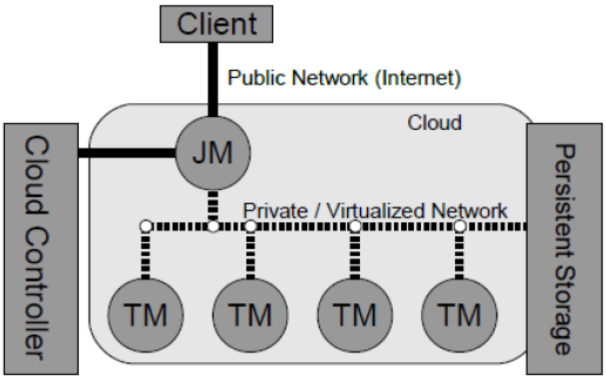

Fig 1. Nephel's Architecture

In this study, they present a novel Turnaround time utility scheduling approach which focuses on both the high priority and the low priority takes that arrive for scheduling. Qian Zhu gives solution of the problem where there is a fixed time-limit as well as a resource budget for a particular task. Within these constraints, an adaptive application needs to maximize the Quality of Service (QoS) metric, more precisely, the value of an application specific benefit function, which captures what is most desirable to compute within the time-limit. They present the design, implementation, and evaluation of a framework that can support adaptive applications in a cloud computing environment. The key component of the framework is a dynamic resource provisioning algorithm, which is based on control theory [6]. Rui Hu, Yong Li includes PaaS platform maintain a low missed deadline ratio and efficient CPU utilization. Thus, compared with those static algorithms, PaaS platform employing FC-LRU can carry more applications and handle more requests [7]. Gihun Jung proposed an adaptive resource allocation model that allocates the consumer's job to an appropriate data centre. The method to adaptively find a proper data centre is based on two evaluations: 1) the geographical distance (network delay) between a consumer and data centres, and 2) the workload of each data centre [8]. Xiaoqiao Men [9] proposed that the Traffic-aware VM Placement Problem (TVMPP) is NP-hard and provide a heuristic algorithm to solve the TVMPP efficiently even for large problem sizes. The proposed algorithm takes a novel two-tier approach: it first partitions VMs and hosts into clusters separately, and then it matches VMs and hosts at cluster level and consequently at individual level. Sujesha Sudevalayam [10] argues the need for network affinity-awareness not only in placement but also in resource provisioning for virtual Machines. First, they empirically quantify the resource savings due to collocation of communicating virtual machines. Then build models based on different resource-usage micro-benchmarks to predict the resource usages when transitioning from non-collocated placements to collocated placements and vice-versa. These resource usage prediction models are usable along-with consolidation and migration procedures to determine requirements of VMs in collocated and non collocated scenarios. Jianhai Chen [11] provides details of all the affinity aware method like data affinity, memory affinity, communication affinity and user affinity among which author works only on communication affinity. So AAGA (Affinity Aware Grouping Allocation) algorithm works to make a group of those machine which is continuously communicate with each other to fulfil the execution of big task and not a single machine in a group is communicate with outsides the group. So MVAG algorithm works in a way to host virtual machine into minimum number of physical machine using vector-bin packing algorithm or any heuristic algorithm.

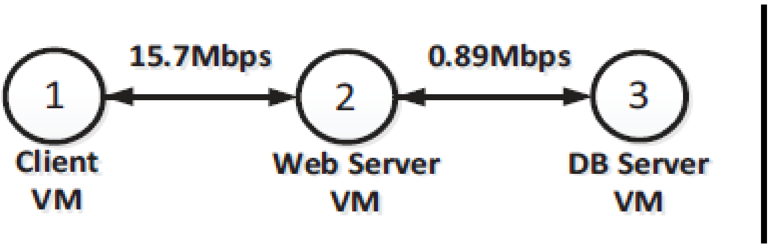

Fig 2: Traffic rate amongst 3 VMs (1–3) running RUBiS benchmark

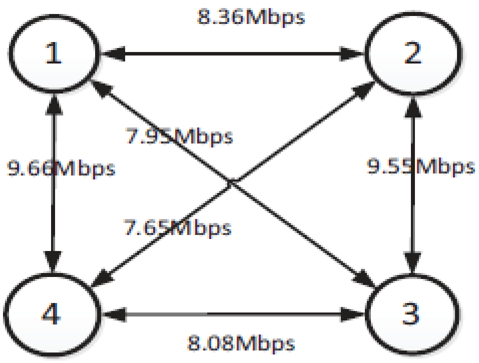

Fig 3: Traffic amongst 4 VMs (1–4) running HPCC benchmark

Communication affinity aware grouping method for collocate the virtual machine into same physical machine or in the same rack or in same data centre. Once grouping is done then algorithm used any technique to host into physical machine like vector bin packing, first fit, best decrease fit etc.

Maximum Virtual machine Allocation by Grouping Algorithm:

Step 1.Input a VM set V with the number N, and a VM-affinity relation set VR with the number E.

Step 2.Initially each VM of V is built as a VM-affinity group set.

Step 3.For each VM in a VM-affinity relation of AR; find the VM-affinity group set which contains the VM,

Step 4.Apply union operation to the two VM-affinity group sets to get a new VM-affinity group set.

Step 5.If there are any other VM-affinity relations, then go to Step 3; otherwise go to Step 6.

Step 6.Output all the VM-affinity group sets as final grouping result.

IV. Problem Definition

As we has seen above different approaches are there for resource allocation but somehow all the techniques which used affinity method to grouping the virtual machine and then host into single physical machine forget one things about resource-contention. We know that grouping of virtual machine gives reduction of network bandwidth as well as time but somehow there should be threshold limit for grouping. Otherwise our motive to make a group is destroying by resource contention.

V. Proposed Work

We want to make an algorithm which makes a group of virtual machine that is communicate with each other but with resource threshold limit. So that at one time grouping is stopped and time consumption is reduce also. We also put some migration techniques to swap the machine when requested job is needed high physical power than they provide by various algorithm. We hope that our Proposed Algorithm will overcome the Existing Problem.

VI. Conclusion

Cloud computing technology is increasingly being used in enterprises and business markets. In cloud paradigm, an effective resource allocation strategy is required for achieving user satisfaction and maximizing the profit for cloud service providers. This paper shows some basic techniques for allocation of Vm and grouping of Vm. And how affinity can be affect to data centre and provider perspective.

References

1. Davide Tammaro, Elias A. Doumith, Sawsan Al Zahr, Jean-Paul Smets, and Maurice Gagnaire "Dynamic Resource Allocation in Cloud Environment Under Time-variant Job Requests" at 2011 Third IEEE International Conference on Coud Computing Technology and Science
2. Chandrashekhar S. Pawar, Rajnikant B. Wagh "Priority Based Dynamic Resource Allocation in Cloud Computing with Modified Waiting Queue" at 2013 International Conference on Intelligent Systems and Signal Processing (ISSP)
3. Jason Sonnek, James Greensky, Robert Reutiman and Abhishek Chandra "Starling: Minimizing Communication Overhead in Virtualized Computing Platforms Using Decentralized Affinity-Aware Migration" at 2010 39th International Conference on Parallel Processing
4. Yi Wei, M. Brian Blake "Adaptive Service Workflow Configuration and Agent-based Virtual Resource Man-agement in the cloud" at 2013 IEEE International Conference on Cloud Engineering
5. Venkatesa Kumar, V. and S. Palaniswami "A Dynamic Resource Allocation Method for Parallel Data Proc-essing in Cloud Computing" in Journal of Computer Science 8 (5): 780-788, 2012 ISSN 1549-3636 © 2012 Science Publications
6. Qian zhu and Gagan Agrawal "Resource Provisioning with Budget Constraints for Adaptive Applications in Cloud Environments" at 1939-1374/11/ © 2011 IEEE
7. Rui Hu, Yong Li, Yan Zhang "Adaptive Resource Management in PaaS Platform Using Feedback Control LRU Algorithm" at 2011 International Conference on Cloud and Service Computing
8. Gihun Jung, Kwang Mong Sim "Agent-based Adaptive Resource Allocation on the Cloud Computing Envi-ronment" in1530-2016/11 © 2011 IEEE
9. Xiaoqiao Meng, Vasileios Pappas, Li Zhang "Improving the Scalability of Data Center Networks With Traf-fic-aware Virtual Machine Placement" at 978-1-4244-5837-0/10 ©2010 IEEE INFOCOM
10. Sujesha Sudevalayam and Purushottam Kulkarni "Affinity-aware Modeling of CPU Usage for Provisioning Virtualized Applications" at 2011 IEEE 4th International Conference on Cloud Computing
11. Jianhai Chen, Kevin Chiew, Deshi Ye, Liangwei Zhu, Wenzhi Chen "AAGA: Affinity-Aware Grouping for Allocation of Virtual Machines" in 2013 IEEE 27th International Conference on Advanced Information Net-working and Applications.

AUTHOR(S) PROFILE

Pratik P. Pandya received the B.E. degree in Computer Engineering in 2010 from V.V.P. Engineering College, Rajkot, Gujart, India. Currently he is pursuing M.Tech. in Computer Engineering from School of Engineering, R.K. University, Rajkot,Gujarat, India. His areas of interest are Cloud Computing.

Hitesh Bheda, recived BE. from Saurashtra University. He has completed his M.E. from L.D. Engg College, Ahmedabad, India. Presently he is working as a Asst.Preofesor in R.K. University, Rajkot, Gujarat, India for 2 years.