AUTOMATIC SCALING OF WEB APPLICATIONS FOR CLOUD COMPUTING SERVICES: A REVIEW

Sushil Deshmukh[1], Sweta Kale[2],

[1] *M.E. Student, Information Technology, RMD Sinhgad School of Engineering, Pune, India*
[2] *Asst. Professor, Information Technology, RMD Sinhgad School of Engineering, Pune, India*

Abstract

Now days there are many web applications can get benefit from automatic scaling property of cloud where the no of resources usage can be scale up and down automatically by cloud service provider. So here present system that provides automatic scaling for web application in cloud environment. So every application instance encapsulated inside virtual machine and model it as the Class Constraint Bin Packing (CCBP) problem. Where each class represents an application and each server is a bin and uses virtualization technology for fault isolation. Now many business customers need good satisfy response services from cloud. So design and develop semi online color set algorithm that achieve good demand satisfaction ratio and as well as when load becomes low it reducing number of server and save energy. Experiment results compare open source implementation of Amazon EC2 demonstrates that system can improve the throughput by 180% over. And system can restore the normal quality of service five times as fast when huge crowd happens. Take supports of green computing to adjusting the placement of application instance adaptively and putting ideal machine into the standby mode.

Key Words: *auto scaling, cloud computing, CCBP, green computing, virtualization etc…*

---***---

1. INTRODUCTION

The elasticity of resource allocation is one of the cited benefits of cloud computing service. As many business customers having scale up and down their application resources usage the customer to buy as many virtual machines instance as they want to operate as like his physical hardware. Now a day cloud provide services as the user still need to decide how much resource are necessary and for how long . so that's is auto-scaling property of cloud where their resources can scale up and down automatically by cloud services will replicate the application which is uploaded by the user on single server onto more and fewer server as the user demands come and goes so user are charged only for what they actually use as called as "pay as you go" facility model.

A simple architecture of cloud computing consist the data centers servers for the web application as well as a switch whose function is balancing the loud and distribute load to set of application server also having set of backend storage server. Switch is typically 7 layer switch whose capture application level information in web request from user and forward to them to application servers with corresponding application running. The switch sometimes runs in redundant pair for fault tolerance. As each server machine can host multiple application so it is important that application should be stateless because every application store their state information in backend storage servers, so that is why they can be replicated safely but it may cause storage servers becomes overloaded but the focus of this work is on application tire presenting a architecture is representative of a large set of internet services hosted in the cloud computing environment even through providing infinite capacity on demand.

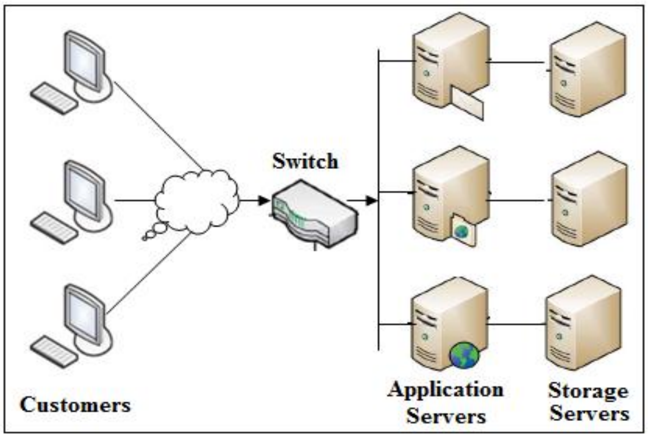

Fig 1: Architecture of web application in cloud computing

The data center capacity in real world is finite but when large number of application access their peck demand around the same time, the available resources in cloud becomes restricted and some of the demand satisfaction ratio or increase percentage of application demand that satisfy successfully. It defines new algorithm and modify CCBP problem and provided higher potential capacity for satisfying application demand. On the other hand when a load or demands of application is low, it is important to save the energy by reducing the number of server used

2. AUTO SCALING IN CLOUD COMPUTING

There are no of way cloud computing provide services these are Software as a service (SaaS), Platform as a service (PaaS), Infrastructure as a service (IaaS)

Mainly the cloud provider is the company that offers the infrastructure and tool for cloud customer to host and maintain their performance of application in cloud. When we talk about scalability and performance management of web application running in software as a service (SaaS) model and platform as a service (PaaS) are the provider responsibility on the other hand when web application running on infrastructure as a service (IaaS) model are completely customer responsibility so when talk about automatic scaling of web application is completely based on SaaS model. In cloud three way of scaling is done horizontal scaling , vertical scaling, auto scaling. Auto scaling is the ability to scale up and scale down the application server's capacity automatically according to customer defines. To maintain the performance when demand is huge it increases the number of instance and decrease automatically when demand reduces to minimize cost.

The auto scaling having number of features

1) It has scale out instances automatically and consistent when demand is increase.
2) It covers unnecessary cloud instances automatically and save money when demand is collapse.
3) It changes the delicate and impassable instances to maintain higher availability of cloud resources of your application.
4) It runs at on demand or spot instance.

3. VIRTUALIZATION

In cloud providing services of web application the virtualization plays important role for fault isolation. It is one of the key enabling technology for cloud computing the main goal of virtualization is to improve the utilization of instance, enable fault tolerance when instance event failure, and easy to dispensation. Virtualization in computer technology is creation of virtual rather than actual. Here it create virtual machine instances for resources allocation. The virtualized resources can be accessed by devices, application, operating system, and by users. Resource virtualization can be categorized into servers, storage, and operating system.

The storage virtualization is allows transparent provisioning storage capacity and simplifies data flexibility and management. The server virtualization is using virtual machine monitor (VVM) layer running between operating system and hardware. The operating system virtualization used abstraction of operating system resource using virtualization layer and that does not runs directly on hardware. This third virtualization implies higher overhead as compare to server virtualization. Due to this it can be used only for application testing but not for production environment. Here for web application of cloud it use server virtualization and also storage virtualization it will use in data centers for fault isolation. When the fault is occur during running process of application servers virtualization pin points the actual cause and location and replace VM instance by creating another VM instance and recovered failed machine quickly.

4. CCBP PROBLEM

It is nothing but the bin packing algorithm solution. There are number of online bin packing algorithm CCBP is one of them it having class constraint bin packing. The class represents application which is divided into number of constraint and the bin is used for the server. Before CCBP there are any fit and next fit. Any fit algorithm can be divided into best-fit, first-fit, worst-fit, Almost-worst-fit. These entire algorithms are used for resource allocation and application placement. The best-fit allocation algorithm place applications on the servers which has smallest block of memory in which it was fit. The idea behind this to use already loaded servers, when possible thus reduce other one for future request and therefore avoid the splitting. But this approach has negative impact on load distribution.

The first-fit allocation algorithm has advantage of using minimum time to selection or detecting the best resource to use. If there are lots of applications that can have request of resources it analyzing all it may take considerable time to choose this algorithm does not search for the best of available servers for the application allocation it chose first one that it finds. so the first-fit algorithm consider server according to order in which they opened and placed each application in the first possible bin.

In worst-fit allocation algorithm it solves the problem that found in previous two algorithms. If there are block of resources to choose particular request requirement it will use best-fit, or first-fit. If block will not match most likely to requested resources perfectly, then after allocation of application very small block of resources are left as unused this block is very small for other request thus it will goes to fragmentation and fragmentation is done in to the CCBP that each application will divided into number of class as constraint and packed it to the servers or bin. In worst-fit it packs every item in the least filled bin.

And the Average-worst-fit is close to worst –fit. If the current application constraint fits in more than one open bin then AWF choose the second least filled bin. Otherwise work like WS. Here the goal is to maximize the demand satisfaction ratio, and minimize the placement change frequency as well as minimize energy consumption. The only difference is the CCBP problem does not solve the "minimize placement change frequency" goal so there it has newly developed modified the CCBP model to give a support for minimizing placement change frequency and it provide new online semi approximation algorithm.

Now in previous CCBP the class constraint represent partial limit of number of application, and capacity of server represent amount of resources available at application server. And here size of item represents an amount of load for particular application. But most online CCBP algorithm does not support for item departure.

Now the key feature of new modified CCBP problem is support for item departure which one is essential to maintain good performance in cloud computing environment where resource demands of web application can change dynamically. Mainly CCBP problem is NP-hard problem there are number of approximation algorithm have been developed. But these entire algorithms assume that the entire input sequence of item is known in advance. But in new environment the demands or request of user can change dynamically and it will change unexpectedly

4. MODIFIED CCBP PROBLEM

Now in modified CCBP mainly focus on the two key point's one is application placement and other one is load distribution. Based on observation develop a semi-online algorithm for CCBP which packs the current item without any knowledge of any sub sequential item in list of input sequence. Here in this scenario color set algorithm is used for the label each class of item with color and arrange them in to color set as per they arrive input sequence.

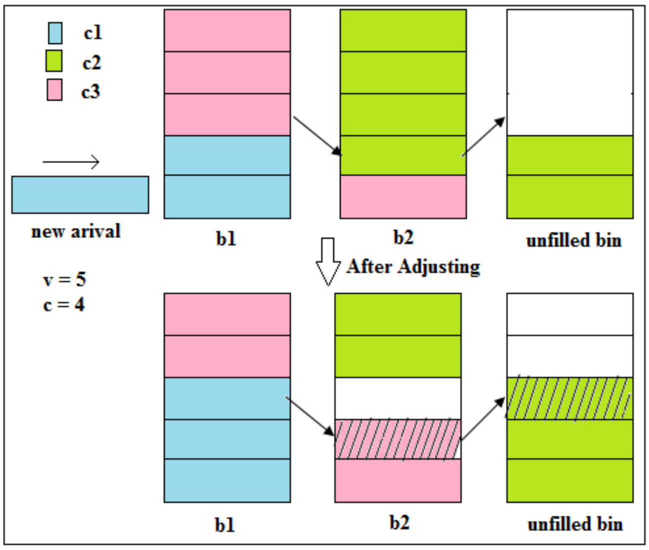

Fig 2 : Modified CCBP at arrival of new item

Item from different color set are packed independently. For packing each item in color set use greedy algorithm it is mathematical process which solve multistep problem by decide which next step will provide most obvious benefit. And item are packed into current bin until the capacity is reached. Here each color set has one unfilled bin so when new item from a specific color set arrives it packed into corresponding unfilled bin.

And if suppose all bin are full with color set then new bin is opened to board the item. Actually here the algorithm attempts to make space for new item in currently full bin by shifting some of item into unfilled bin. If the application load increase as the arrival of new item with the corresponding color, if the unfilled bin does not exist that color already then new color is to add in to bin as shown in fig 2. These movement of item is hypothetical and used only to calculate new load distribution the shifting of item one by one is make chain if we cannot find such a chain, the new color is add to unfilled bin which is starting new application instance if color set has no unfilled bin then new bin allocated.

Now the application load decrease is modeled as the departure of previously packed items.

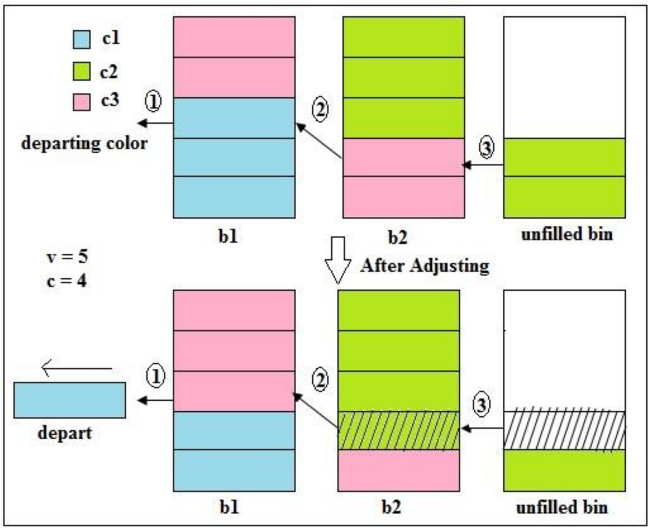

Fig 3 : Modified CCBP at departure of item

Main note is departure event is associated with number of specific color, not with specific item. Main advantage is the algorithm has freedom to select which item of particular color is to remove. And challenge is to maintain property that every color set has one unfilled bin departure working as follows.

- If in the color set does not present unfilled bin then remove any item of that color and the output bin becomes unfilled bin
- If the unfilled bin contains the departing color, so corresponding item removed directly
- In this case if need to remove an item from currently full bin then fill the hole with an item in form somewhere else

And finally last item of particular color leaves that selected color can be removed from its color set this is nothing but the closing down the last instance of an application when load reduce to zero. And color set become unfilled the challenge is here is to maintain property that there is at most one unfilled color set in the system.

5. SUMMARY

- In these sections first summarized the automatic scaling problem in the cloud computing for web application and model it as the CCBP class constraint bin packing problem to solve the problem it compared with existing bin packing solution as well as creatively support for item departure that effectively avoids frequent placement change
- To give a way support from green computing and greedy algorithm that can adjusting the placement of application instance adaptively and put ideal virtual machine in standby mode.
- This modified algorithm is highly efficient and scalable which is achieving high demand satisfaction ratio.
- When compare the performance of this system with open source implementation of the Amazon EC2

that it can restore normal quality of service fast at the huge number of crowed is arrived.

- This modified CCBP having fast restart technique based on virtualization of virtual machine.

6. CONCLUSIONS

The presenting design and implementation of system which is scale up and scale down number of application instance automatically based on user demand. It develop color set algorithm and also use greedy algorithm to decide the application placement and load distribution. This reaches high satisfaction ratio of an application demand even when the load becomes very high, also saves the energy by reducing number of running instance of virtual machine when the load is low. It use virtualization for fault isolation an maintain good health of cloud application server.

ACKNOWLEDGEMENT

I sincerely thank to Sweta Kale for her continuous and constructive support. I would also like to thank Dhara Kurian for encourage the work on cloud computing services. I would also like to thank Sweta Kale her helpful comments on paper

REFERENCES

[1] Zhen Xiao, Senior Member, IEEE, Qi Chen, and Haipeng Luo ”Automatic Scaling of Internet Applications for Cloud Computing Services” IEEE TRANSACTIONS ON COMPUTERS, VOL. 63, NO. 5, MAY 2014

[2] Leah Epstein, Lene M. Favrholdt, and Jens S. Kohrt “Comparing Online Algorithms for Bin Packing Problems”

[3] M.Kriushanth, L. Arockiam and G. Justy Mirobi”Auto Scaling in Cloud Computing: An Overview” International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 7, July 2013

[4] Tania Lorido-Botr´an, Jos´e Miguel-Alonso, Jos´e A. Lozano Auto-scaling Techniques for Elastic Applications in Cloud Environments September 5, 2012

[5] C. Chekuri and S. Khanna, “On multidimensional packing problems,” SIAM J. Comput., vol. 33, no. 1, pp. 837–851, 2004.

[6] H. Shachnai and T. Tamir, “Tight bounds for online class constrained packing,” Theor. Comput. Sci., vol. 321, no. 1, pp. 103–123, 2004

BIOGRAPHIES

Sushil Raghunath Deshmukh has completed bachelor degree from Dr. Babasaheb Ambedkar Marathwada University and now M.E Student at RMD Sinhgad school of Engineering, Pune, Maharashtra, India

Sweta Kale, Dept. of Information Technology, Working at RMD Sinhgad school of Engineering, Warje Savitribai Phule, Pune university ,Maharashtra, India