Predicting Future Resource Requirement for Efficient Resource Management in Cloud

B.V.V.S Prasad
M.C.A., M.Tech,(Ph.D.)
Associative Prof of Malineni Lakshmaiah Women's Engineering College, Guntur, Andhra Pradesh, India

Sheba Angel
M.Tech Research Scholar of Malineni Lakshmaiah Women's Engineering College, Guntur, Andhra Pradesh, India

ABSTRACT

Cloud Computing became an optimal solution for business customers to maintain and promote their business needs to clients via cloud services like IaaS, SaaS, and PaaS. On-Demand and pay-per-use scale up methodologies attracted organizations for cloud adoption and migration. Due to the increased demand for cloud services from users, Efficient Resource Management in cloud computing become an important task. In order to achieve resource multiplexing in cloud computing, recent researches were introduced dynamic resource allocation through virtual machines. Existing dynamic approaches followed un-evenness procedures to allocate the available resources based on current workload of systems. Unexpected demand for huge amount of resources in future may cause allocation failure or system hang problem. In this paper we present a new systematic approach to predict the future resource demands of cloud from past usage. This approach analyzes the resource allocation logs of virtual server, SLA agreements and follows the resource prediction algorithm to estimate future needs to avoid allocation failure problem in cloud resource management. Experimental results are supporting our strategy is more scalable and reliable than existing approaches.

Keywords

Dynamic resource allocation, Cloud computing, Resource Prediction Algorithm, Virtual Cloud Servers

1. INTRODUCTION

Cloud computing is an emerging technology to provide low cost, high end, reliable QoS (Quality of Service) to customers as IaaS, PaaS, SaaS and others. Cloud services offering on-demand, pay-per-use scale up methodologies to business customers to alleviate the infra management burden on them. Majority of Business customers interested towards cloud computing and they started their app migration with cloud environment to promote their business operations to end client with low investments and high availability. Due to this increased adoption, Resource Management in Cloud (RMC) becomes an important research aspect in this area. Earlier approaches [1, 3] were used evenness procedure in resource distribution to allocate the available resources among the running applications. This approach may leads to resource over flow due to high amount of resource allocation than required and resource underflow due to less amount of resource allocation than required. Always resource needs for a running application changes from time to time depends on number of live clients.

In order to overcome resource overflow and resource underflow problems from evenness distribution recent researches were introduced dynamic resource management [4 and 5] with virtual systems. These systems will consider the available resources at server and allocates them to applications based on application workload requirements. To achieve this dynamic managements systems follows unevenness algorithms and on demand resource allocation strategies. This approach will manage the resources dynamically in an efficient manner with virtualization of cloud systems. Dynamic mapping of virtual requirements with physical resources will also help to avoid SLA violations [6] in cloud environment. Sometimes unexpected demand for huge amount of resources in future may cause allocation failure or system hang problem.

In order to mitigate these problems, in this paper we present a new systematic approach to predict the future resource demands of cloud from past usage. This approach analyzes the resource allocation logs of virtual server, SLA agreements and follows the demand prediction algorithm to estimate future needs to avoid allocation failure problem in cloud resource management. Our approach uses the present and past statistics to predict the future requirements in an efficient manner. To do this we proposed two different methodologies in this paper are (i) hours-bounded (ii) days-bounded resource prediction techniques. By integrating the results of these methodologies our approach assess the reliable resource requirements in future. Experimental results are supporting our strategy is more scalable and reliable than existing approaches.

The rest of the paper is organized as related work explored in section 2, our future resource demand prediction and prediction techniques in section 3, experiments and results from section 4, conclusion from section 5 and references from section 6.

2. RELATED WORK

Increasing demand for cloud migration enforces the efficient resource management in cloud computing. Static resource allocation and dynamic resource allocation were the two important methodologies in this area for resource management. In this section we explore the detail description of these two methodologies.

2.1 Static Resource Allocation:

In this method, all systems [1, 3 and 9] will share the available resources (CPU time, Memory, Registers and Network bandwidth etc.) equally without having any priorities. For example a cloud server with set of resources R should be equally shared among a list of processes P like ∀ p1 to pn where pi ∈ P, ∀ r1 to rn where ri ∈ R and any ri equals to rj and ri == rj will be always true. Priority cannot be assessed and maintained in this approach based on any parameters. Simply these approaches are following evenness in resource allocation which may cause resource overflow and resource underflow errors. This approach does not support Green Computing [8] because of having resource overflow problem. Multiplexing the resources in cloud computing is very important because of increasing cloud adoptions and unstable user communications.

2.2 Dynamic Resource Allocation:

Recently some authors proposed dynamic resource management [4, 5] in their research papers for efficient resource allocation based workload through virtualization. Xiao and song et al introduced unevenness procedure to distribute physical resources among virtual servers of cloud environment. This approach will be automatically refreshed eventually to update the allocated resources of server based on the workload. In order to predict the hot spot and cold spots [7, 10] this method will determine the threshold from periodic evaluations of workload. CPU and Network resources are monitored by scheduling algorithms and memory resources are by swap activities in this area. While monitoring the requirements at run time, dynamic resource management predicts the workload of applications and verifies the hot spot and cold spot threshold to distribute the available resources among physical systems. After analysis they sort the applications based on hot spot and cold spot values to identify the priority for resource allocation. In this case the high level hot spot system will be migrated with high level cold spot system to acquire the additional resources to manage the workload. This is proven as an efficient way manages the resources in cloud environment to mitigate hot spot (resource underflow) and cold spot (resource overflow) problems.

3. PREDICTING FUTURE RESOURCE REQUIREMENT

Recent cloud architectures are facing resource management problem, due to unexpected requirement of huge resources (CPU time, memory, networks etc.) in an asynchronous manner. Current dynamic resource allocation methodologies [5, 11] are having the capability to map virtual system resources with physical systems dynamically depends on work load. This process will adjust the available resources with the help of hotspot and cold spot migration among the physical machines. Dynamic resource allocation may fail under some circumstances, when suddenly there is an unexpected huge resource requirement for a physical machine.

To address the above problem, in this paper we introduced “Predicting Future Resource Requirement Framework (PFRRF)” to assess the future resource needs. This framework is an extension to dynamic resource allocation and management architecture. PFRRF considers the log files of resource allocation system to analyze and estimate future requirements. To achieve this our framework uses the Resource Prediction Algorithm(RPA) which takes structured log file content as training data and time bounded methodologies for decision making process in resource assessing. Semi-structured log file data would be transformed to structured log files, which are having the periodic resource allocation charts (PRAC) for every physical machine running on cloud environment. PRAC contains individual resource level mapping to each physical machine and hot spot, cold spot thresholds. After that PRAC’s of every physical system are sorted on date parameter and clustered individually at physical system level. These individual clusters are given as input data to RPA to predict the future workload of every physical machine. After computing of physical machines prediction results, resource migration will be performed based on resource overflow and under flow results. This migration helps to know the additional resource requirement for future based on current and past log requirements. At the end, sum of every physical machine level additional requirement will be the final requirements, which to be added to cloud resource pool to avoid future underflows and hanging problems. This framework architecture is presented in fig.1 as described in above.

4. PREDICTING FUTURE RESOURCE REQUIREMENT

Recent cloud architectures are facing resource management problem, due to unexpected requirement of huge resources (CPU time, memory, networks etc.) in an asynchronous manner. Current dynamic resource allocation methodologies [5, 11] are having the capability to map virtual system resources with physical systems dynamically depends on work load.

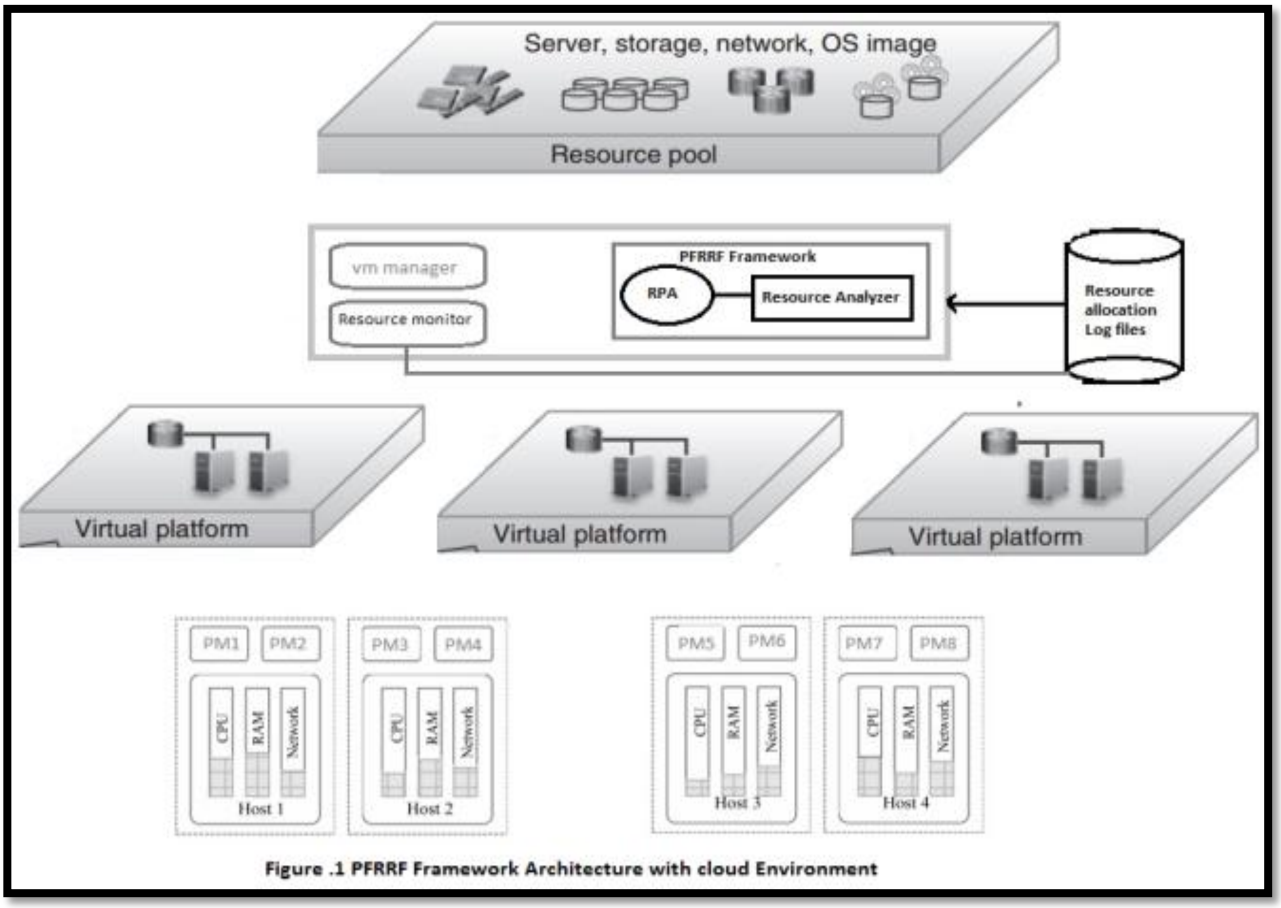

Figure .1 PFRRF Framework Architecture with cloud Environment

This process will adjust the available resources with the help of hotspot and cold spot migration among the physical machines. Dynamic resource allocation may fail under some circumstances, when suddenly there is an unexpected huge resource requirement for a physical machine. To address the above problem, in this paper we introduced “Predicting Future Resource Requirement Framework (PFRRF)” to assess the future resource needs. This framework is an extension to dynamic resource allocation and management architecture. PFRRF considers the log files of resource allocation system to analyze and estimate future requirements. To achieve this our framework uses the Resource Prediction Algorithm(RPA) which takes structured log file content as training data and time bounded methodologies for decision making process in resource assessing. Semi-structured log file data would be transformed to structured log files, which are having the periodic resource allocation charts (PRAC) for every physical machine running on cloud environment. PRAC contains individual resource level mapping to each physical machine and hot spot, cold spot thresholds. After that PRAC’s of every physical system are sorted on date parameter and clustered individually at physical system level. These individual clusters are given as input data to RPA to predict the future workload of every physical machine. After computing of physical machines prediction results, resource migration will be performed based on resource overflow and under flow results. This migration helps to know the additional resource requirement for future based on current and past log requirements. At the end, sum of every physical machine level additional requirement will be the final requirements, which to be added to cloud resource pool to avoid future underflows and hanging problems. This framework architecture is presented in fig.1 as described in above.

5. TIME BOUNDED RESOURCE PREDICTION

In this paper we are introducing the naïve concept is time bounded resource prediction for future. After analyzing many existing resource allocation methodologies we observed that resource allocation requirements are changing from time to time, here the time stands for hour to hour and day to day. Henceforth predicting resource requirements for future will be a time bound operation to support the green computing policy specifications. To avoid the complexity of time bounded prediction we are analyzing from hour to day. While considering accuracy and effectiveness in resource prediction the time bounded prediction is having the high scalability. To achieve this we proposed three different time bounded resource prediction methodologies.

5.1 Hour bounded Resource Prediction: From the observations of many web applications we found that some applications needs the huge resources in a day at specific time hours. For example if we consider IRCTC (India’s biggest online train reservation system) needs approximately three times more amount of memory, network and process resources during “Tatkal Reservation” on every day between 10:00 AM to 12:00 PM (2 hrs.). If we calculated the workload at day level in this case, results the inaccurate and unreliable estimations. To avoid this, our framework uses the hour wise predictions and prepares the resource prediction chart at hour level. RN, RP, RM are respectively the sets of network, process and memory resources and TH is a set of hour time elements range from th1 to th24. For any given time hour thi we can calculate the workload like shown below:

Workload for a system sj is $W(thi) = \alpha * (O(RN) + \partial(RN) + {}_{\beta}(RM))$.

5.2 Day bounded Resource Prediction:

Generally majority of the service oriented applications like internet service providers, online gaming websites, online shopping websites will be busy on weekends and holidays. Instead of at specific hours of a day they will be busy on complete day itself. In this case, hour level calculation prediction is not so efficient and more complex; henceforth we introduced the day level resource prediction. This will consider all days of a month and calculates the day wise workload similar to hour bounded resource prediction. Here we have to notice that day bounded prediction will be subject to expected holidays and unexpected holidays of a month.

5.3 Resource Prediction Algorithm (RPA):

After preparation of Time bounded Prediction, RPA takes the resource prediction chart as input and process the chart data to compare against the available resource to predict the additional resource usage in future as shown below.

Resource Prediction Algorithm (RPA):

```
Input: present & past resource usage chart
output: future resource prediction chart(FRPC)
PM : Physical Machine,  RUC : resource usage chart
foreach pm in cloud do
RUC ← getUsageChart(pm)
SRUC ← doUsageAnalysis(RUC)
PRUC ← predictFutureRequirements(SRUC)
   if(PRUC <= THRESOLD) setOverFlowFlag()
   else setUnderFlowFlag()
wishList.addToWishlist(PRUC)
end foreach;
FRPC = doMigration(wishList)
return FRPC
End
```

First RPA considers every physical machine in cloud and generates the resource usage chart (RUC) for every machine based on cloud resource consuming log file. Based on RUC data RPA creates the SRUC to predict future requirements PRUC. If the requirements are less than the threshold the resource underflow flag will be set else resource overflow flag will be set and be added to wish list. Migration method[12] will take the wish list as input and performs mapping to return FRPC.

6. EXPERIMENTS

In this section we discuss about the performance of our framework and comparison with other approaches to prove the efficiency. Our experiments mainly concentrated on prediction of future resource requirements based on present and past usage details from server log files. In this case we observed the three important resources are memory, process and network resource. In order to perform these experiments we were selected the Unix based private cloud hosting center which is managing more than 20 applications of various technologies. On pilot basis we were taken 8 applications to adopt our framework process separately. For the training data we had observed the last 12 months resource allocation charts from log files. This cloud center is already following its own resource allocation technology for all physical machines running on hosting center. Along with allocation accuracy checking this testing also considers the SLA violations with every physical machine.

RPA evaluation: Every physical Machine of Cloud hosting center is having 8 GB of RAM, Core i7 (2nd Gen) Processor and 1 TB of Hard Disk. Apart from this additional resources are available with the hosting center to allocate dynamically as per the requirements. We deploy 8 virtual machines to map with 8 physical machines which are running the applications. For every five minutes the resource allocation will be updated to adjust the resources among physical machines and writes the same on log files. We had given the last one year resource allocation chart to experimental cloud center. After analyzing this data, our framework predicted the next one month resource requirements hour wise and day wise as per applicability at physical machine level. We had compared the framework predicted resources for future with the consumed resources on the specific hours and days. These experimental results are shown in the below table .1 and accuracy of prediction also represented with graphs.

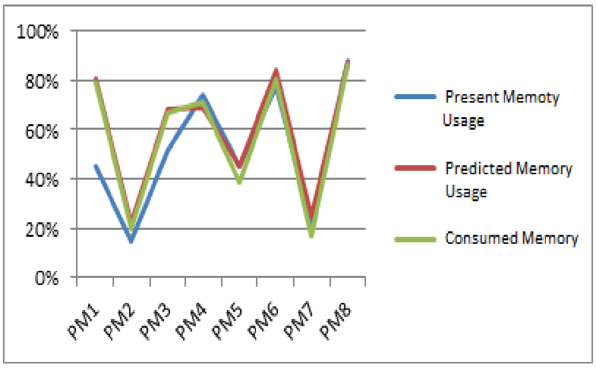

Fig .2 Predicted vs consumed Memory utilization representation

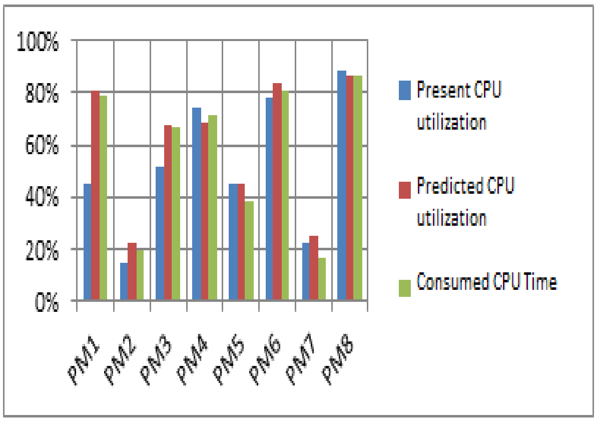

Fig .3 Predicted vs. consumed CPU Time utilization representation

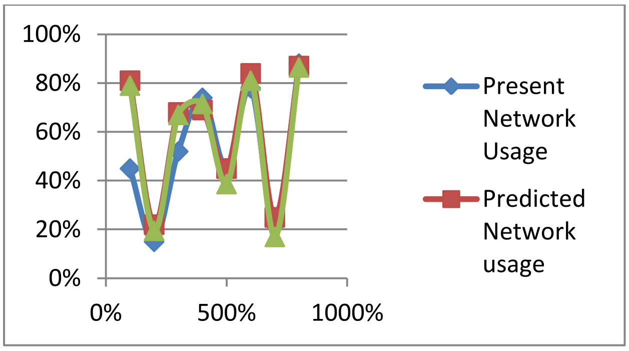

Fig .4 Predicted vs consumed Network utilization representation

7. CONCLUSION

In this paper, we presented Predicting Future Resource Requirement Framework (PFRRF) to assess the future resource needs. This framework is an extension to dynamic resource allocation and management architecture. Our system predicts the future resource needs based on present and past allocation data from resource log files. We use the Resource Prediction Algorithm (RPA) to assess the future need effectively. This approach achieved high accuracy in the area of predicting the future resource needs. We proposed hour bounded and day bounded resource prediction methodologies depends on the applicability. This research helps to avoid resource underflow and overflow problems efficiently. Experiments are proving the prediction accuracy and green computing for memory, CPU time and network resources in physical machines of cloud.

8. REFERENCES

[1] R. Nathuji and K. Schwan, “Virtualpower: coordinated power management in virtualized enterprise systems,” in Proc. of the ACM SIGOPS symposium on Operating systems principles (SOSP’07), 2007.

[2] S. Srikantaiah, A. Kansal, and F. Zhao, “Energy aware consolidation for cloud computing,”Cluster Computing, vol. 12,pp. 1–15, 2009.

[3] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat,and R. P. Doyle, “Managing energy and server resources inhosting centers,” ACM New York, NY, USA,2001, pp. 103–116.

[4] Atsuo Inomata, TaikiMorikawa, Minoru Ikebe, Sk.Md. Mizanur Rahman: Proposal and Evaluation of Dynamin Resource Allocation Method Based on the Load Of VMs on IaaS (IEEE,2010),978-1-4244-8704-2/11.

[5] etahiWuhib and Rolf Stadler : Distributed monitoring and resource management for Large cloud environments(IEEE,2011),pp.970-975.

[6] Hadi Goudaezi and Massoud Pedram: Multidimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems IEEE 4th International conference on Cloud computing 2011,pp.324-331.

[7] Wei-Yu Lin et al. : Dynamic Auction Mechanism for Cloud Resource Allocation: 2010 IEEE/ACM 10th International Conference on Cluster, Cloud and Grid Computing, pp.591-592

[8] Rerngvit Yanggratoke, Fetahi Wuhib and Rolf Stadler: Gossip-based resource allocation for green computing in Large Clouds: 7th International conference on network and service management, Paris, France, 24-28 October, 2011.D. Meisner, B. Gold, T. Wenisch, Powernap: eliminating server idle power, ACM SIGPLAN Notices 44 (3) (2009) 205–216.

[9] Yang wt.al A profile based approach to Just in time scalability for cloud applications, IEEE international conference on cloud computing ,2009,pp 9-16.

[10] C. Tang, M. Steinder, M. Spreitzer, and G. Pacifici, “A scalable application placement controller for enterprise data centers,” in Proc. Of the International World Wide Web Conference (WWW’07), May 2007.

[11] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach,I. Pratt, and A. Warfield, “Live migration of virtual machines,” in Proc.of the Symposium on Networked Systems Design and Implementation (NSDI’05), May 2005.