Volume 2, No. 3, May-June 2011

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Survey on Adaptive Resource Allocation for Multi-tier Web application using Ad-hoc Clouds

R.Rajeshkannan
,Asst.Professor,
School of Computing Science and Engineering
VIT University,Vellore
rajeshkannan.r@vit.ac.in

Abstract: Ad-hoc Cloud Computing allows us to abstract distributed expandable IT resources behind an interface that promotes scalability and dynamic resource allocation. The prototype exploits adaptive allocation of cloud resources to scale gracefully in the presence of rapid increases in workload. The objective of this research is to investigate creating a framework incorporating web services and ad-hoc clouds harvesting unused computing resources in a non intrusive manner. The underused computing resources within a participating enterprise represent a major untapped computing resource and can be utilized to improve the quality of service. The harvested resources can be made available in the form of ad-hoc clouds which can be configured dynamically based on the resource availability and service demands. This research proposes extending the concept of the cloud to encompass not only server-farm resources but all resources accessible by the user. This brings the resources of the home PC and personal mobile devices in to the cloud and promotes the deployment of highly-distributed component based applications with fat user interfaces. For the end user, a web application deployed on a cloud is presented no differently to a web application deployed on a stand-alone web server. This model works well for web applications.

Key words: Cloud computing, Resource allocation, Service requirements.

I. INTRODUCTION

An ad-hoc cloud computing is a combination of physically and virtually connected resources. In this ad hoc cloud computing offering platform as a service which facilitate for application design, application development, testing, deployment and hosting as well as application services such as web service integration, security, scalability and storage. The platform based service provided by a company, group, community, or government that provides a platform in which to develop software applications, usually web based, with immediate abstractions of the underlying infrastructure. The Ad-hoc cloud service provider provides quality attributes for multi-tier web application. The reason is network traffic is highly unpredictable and response time also depends to other factor, so allocate the resource dynamically based on traffic grows. By using resources on demand in ad-hoc cloud data centers, developers can extensively reduce resource management and deployments. Resources are available as virtual machines that appear as traditional physical machines. In this paper, we will discuss the challenges of adaptive resource allocation in platform based system (PBS), and the current state of the art in handling dynamic resource allocation for various computing and network systems which are useful for dynamic resource allocation in PBS. We will then present an approach to adaptively allocating the system resources of servers to their services in runtime to satisfy one of the most important QoS requirements, the throughput, of multiple workflows in PBS.

II. RELATED WORK

This section discussed some past work concerning capacity planning of IT resource that bears some relationship to in this work. Beyond the capacity planning process, many papers investigate cloud-based solutions for dynamic

provisioning of IT capacity. It is interesting to notice that, as far as we know, none of them have considered an IT infrastructure entirely composed by ad-hoc cloud computing resources.

Recently, Maciel Jr. et al. have proposed a hybrid dynamic IT infrastructure [1]. In that hybrid system, computing power can be obtained from in-house dedicated resources, via resources acquired from cloud computing providers, and resources received from a best-effort peer-to-peer (P2P) grid.

A dynamic infrastructure was also proposed by Rich and Altmann [2]. This paper discussed in house resources and short term and long term planning of IT infrastructure. In this infrastructure only dedicated in house machines. It is more difficult to address in dynamic environment. Assuncao et al. [3] inspect the advantage of using cloud computing to increase the capacity of local IT infrastructures. This paper evaluates resource utilization in order to reduce the response time of user request and demonstrate that scheduling strategies can reduce the overall cost.

Popovici and Wilkes [4] delivered the idea an economics-oriented approach for a service provider, how to

give the priority of customer request and extend in this works by Chun and Culler [14] and Irwin et al. [5], by considering job based services to its user and rents resources from a resource provider. It is more difficult to select the requests and assume that the service provider will have some uncertainty about the availability of the resources necessary to fulfill the requests. Popovici and Wilkes defined risk-aware heuristics for admission control and scheduling. They were aimed at maximizing the resource availability and the authors do not account for a previous capacity planning process in order to address the problem of resource availability.

There was a wide range of literature on on-demand resource allocation for Web applications [6], [7], [8], [9]. These research groups are, in general, not concerned with business metrics and they do not take providers costs into account. Their goal is to find out the amount of resources to be provisioned to an application during its execution. More important, they focus on the application scheduling. They study the dynamic provisioning problem including the selection of the most cost-effective PaaS providers.

III. COMPARISON OF DIFFERENT RESOURCE ALLOCATION METRICS

Walsh et al. [10], recommend two-layer architecture and it used in the dynamic resource management. This architecture encompass of local agents, application agents, global decision module and etc. The local and application agents are tightly coupled with each Application environment. A global decision module computes configurations of the entire data center in a centralized manner. The local agents calculate utilities of each Application environment then the results are sent to the global decision module. The global decision module is taken a responsible for computing configuration of resources given the utilities declared by the local agents. The main aim is to assign enough resources to each Application Environment without violating SLAs while not exceeding the total resources in the data center. The global decision module computes new configurations either at the end of fixed control intervals or in an event triggered manner where events are considered to be current or anticipated SLA violations.

Studies outlined in Bennani et al. [11], Chess et al. [12], Tesauro [13], Tesauro et al. [14], [15], and Das et al. [9] adopted non-virtualized data centers approach. They are focused on performance modeling of Application Environments, leveraging forecasting methods based on different analytical models. Bennani and Menasce [11] investigated on a queuing theoretic approach that to be solved performance modeling problem, Tesauro [13],Tesauro et al. [14], [8] considered a pure decompositional reinforcement learning approach and a hybridization with the queuing theoretic approach. Chess et al. and Das et al. [12], [16], concentrated the same architecture and utility model and used to build a commercialized computing system.

Finally this research has focused on virtualized data centers where the components of Application Environments are encapsulated in Virtual Machines. Initially Almeida et al. [17], worked data center utilization and it has been explicitly considered as a major criterion in dynamic and autonomous resource management. In addition, this work sketched data center utilization as a major factor in both short term and long-term resource planning. Other research outlined in Khanna et al. [18], Bobroff et al. [19], Wood et al. [20], Wang et al. [14], Kochut [21], Hermenier et al. [3], and Van and Tran [22] have also adopted a centralized configuration as outlined in [1]. In some studies the cost of migrations—in terms of overhead—during configurations was taken into account.

Hermenier et al. [23] deals on reducing the number of migrations by using constraint solving method. The general method is to find a set of possible configurations and finding the most suitable one that maximizes global utility and minimizes the number of migrations. This same method is also adopted in the follow-up work of Van and Tran [16]. Although these methods are very efficient in relatively small data centers, we believe that they can potentially suffer from scalability, feasibility and flexibility issues.

Villela et al. [24] suggested the optimal resource allocation to the application server tier in a multi-tier e-commerce Web application. This application hosting environment will increase the profit of service provider. It identified the request arrival process as a Poisson process by analyzing the traces of a real Web application and formalized the resource allocation problem to maximize the profit of service providers by optimal resource allocation to the application server tier. They present and evaluate three different approximation methods for optimal resource allocation using simulations.

Dubeyb et al. [25] present the initial results of dynamic regression and queuing modeling techniques to obtain the approximate system performance model for multi-tier Web application hosting on virtualized data centers. They use the dayTrader [9] benchmark Web application to evaluate their prototype system on a Xen-based virtualization platform. Jungy et al. [1] present off-line techniques to generate adaptation policies for multi-tier applications hosted on virtualized data center. The purpose of an adaptation policy is to provide optimal configurations of an application for the given workload. They present a queuing model with optimization techniques to generate optimal system configuration for a multi-tier application. Their model is able to identify the number of replicas for tiers, tiers placement, and CPU allocation for tiers.

IV.CONCLUSION

This paper argue the problem of autonomic virtual resource management for hosting service platforms with a two-level and multi level architecture which isolates application specific functions from a generic decision-making layer. We take into account both high-level performance goals of the hosted applications and objectives related to the placement of virtual machines on physical machines. Self optimization is achieved through a combination of utility functions and a constraint programming approach. Utility functions provide a high-level way to express and quantify application satisfaction with regard to SLA and to trade-off between multiple objectives which might conflict with one another Simulation experiments have been conducted to

validate our architecture and algorithms. The autonomic management system is being designed a component-based framework with a clear separation between generic mechanisms and pluggable modules.

V. REFERENCE

[1] P. D. Maciel Jr., F. Figueiredo, D. Candeia, F. Brasileiro, and A. Coˆelho, "On the Planning of a Hybrid IT Infrastructure," IEEE/IFIP Network & Operations Management Symposium (NOMS'08), April 2008.

[2] M. Risch and J. Altmann, "Capacity planning in economic grid markets." in GPC, ser. Lecture Notes in Computer Science, N. Abdennadher and D. Petcu, Eds., vol. 5529. Springer, 2009, pp. 1–12. [Online]. Available: http://dblp.uni-trier.de/db/conf/gpc/gpc2009.html#RischA09

[3] M. D. de Assuncao, A. di Costanzo, and R. Buyya, "Evaluating the cost benefit of using cloud computing to extend the capacity of clusters," inHPDC '09: Proceedings of the 18th ACM international symposium on high performance distributed computing. New York, NY, USA: ACM,2009, pp. 141–150.

[4] F. I. Popovici and J. Wilkes, "Profitable services in an uncertain world,"in Proceedings of the 2005 ACM/IEEE Conference on Supercomputing (SC'05). Washington, DC, USA: IEEE Computer Society, 2005, p. 36.

[5]. B. N. Chun and D. E. Culler, "User-centric performance analysis of market-based cluster batch schedulers," in Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID'02). Washington, DC, USA: IEEE Computer Society, 2002,p. 30.

[6] S. Ranjan, J. Rolia, H. Fu, and E. Knightly, "Qos-driven server migration for internet data centers," in Proceedings of the International Workshopon Quality of Service, Miami, Florida, EUA, Maio 2002, pp. 3 – 12.

[7] E. Lassettre, D. W. Coleman, Y. Diao, S. Froehlich, J. L. Hellerstein,L. Hsiung, T. Mummert, M. Raghavachari, G. Parker, L. Russell,M. Surendra, V. Tseng, N. Wadia, and P. Ye, "Dynamic surge protection:An approach to handling unexpected workload surges with resource actions that have dead times," in 14th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management, ser. Lecture Notes in Computer Science, vol. 2867. Springer, 2003, pp. 82–92.

[8] B. Urgaonkar, P. Shenoy, A. Chandra, P. Goyal, and T. Wood, "Agile dynamic provisioning of multi-tier internet applications," ACM Trans.Auton. Adapt. Syst., vol. 3, no. 1, pp. 1–39, 2008.

[9] T. Setzer, A. Stage, and M. Bichler, "Automated capacity management and selection of infrastructure-as-a-service providers," in Proceedings of the 4th Workshop on Business-Driven IT Management, New York, USA,2009, pp. 20–23.

[10] W. E. Walsh, G. Tesauro, J. O. Kephart, and R. Das, "Utility Functions in Autonomic Systems," in ICAC '04: Proceedings of the First International Conference on Autonomic Computing. IEEE Computer Society, 2004, pp. 70–77.

[11] M. N. Bennani and D. A. Menasce, "Resource Allocation for Autonomic Data Centers using Analytic Performance Models," in ICAC'05: Proceedings of the Second International Conference on AutomaticComputing, 2005, pp. 229–240.

[12] D. Chess, A. Segal, I. Whalley, and S. White, "Unity: Experienceswith a Prototype Autonomic Computing System," in 2004. Proceedings. International Conference on Autonomic Computing, 2004, pp. 140–147

[13] G. Tesauro, "Online Resource Allocation Using Decompositional Reinforcement Learning," in AAAI'05: Proceedings of the 20th National Conference on Artificial Intelligence. AAAI Press, 2005, pp. 886–891.

[14] G. Tesauro, R. Das, W. Walsh, and J. Kephart, "Utility-Function-Driven Resource Allocation in Autonomic Systems," in Autonomic Computing,2005. ICAC 2005. Proceedings. Second International Conference on,2005, pp. 342–343.

[15] G. Tesauro, N. Jong, R. Das, and M. Bennani, "A Hybrid ReinforcementLearning Approach to Autonomic Resource Allocation," in ICAC '06:Proceedings of the 2006 IEEE International Conference on AutonomicComputing. IEEE Computer Society, 2006, pp. 65–73.

[16] R. Das, J. Kephart, I. Whalley, and P. Vytas, "Towards Commercialization of Utility-based Resource Allocation," in ICAC '06: IEEE International Conference on Autonomic Computing, 2006, pp. 287–290.

[17] J. Almeida, V. Almeida, D. Ardagna, C. Francalanci, and M. Trubian, "Resource Management in the Autonomic Service-Oriented Architecture,"in ICAC '06: IEEE International Conference on Autonomic Computing, 2006, pp. 84–92.

[18] G. Khanna, K. Beaty, G. Kar, and A. Kochut, "Application Performance Management in Virtualized Server Environments," in Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP, 2006, pp. 373–381.

[19] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic Placement of Virtual Machines for Managing SLA Violations," in IM '07: 10th IFIP/IEEE International Symposium on Integrated Network Management, 2007, pp. 119–128.

[20] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Black-Box and Gray-Box Strategies for Virtual Machine Migration," in 4th USENIX Symposium on Networked Systems Design and Implementation, 2007,pp. 229–242.

[21] A. Kochut, "On Impact of Dynamic Virtual Machine Reallocation on Data Center Efficiency," 2008, pp. 1 –8.

[21] X. Wang, D. Lan, G. Wang, X. Fang, M. Ye, Y. Chen, and Q. Wang, "Appliance-Based Autonomic Provisioning Framework for Virtualized Outsourcing Data Center," in Autonomic Computing, 2007. ICAC '07. Fourth International Conference on, 2007, pp. 29–29.

[22] H. N. Van and F. D. Tran, "Autonomic Virtual Resource Management for Service Hosting Platforms," in ICES '09: Proceedings of the International Conference on Software Enginnering Workshop on Software Engineering Challenges of Cloud Computing. IEEE Computer Society, 2009, pp. 1–8.

[23] F. Hermenier, X. Lorca, J. M. Menaud, G. Muller, and J. Lawall, "Entropy:A Consolidation Manager for Clusters," in VEE '09: Proceeding of the 2009 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, 2009, pp. 41–50.

[24] D. Villela, P. Pradhan, and D. Rubenstein, "Provisioning servers in the application tier for e-commerce systems,"

ACM Transaction on Internet Technology, vol. 7, no. 1, p. 7, 2007.

[25].A. Dubey, R. Mehrotra, S. Abdelwahed, and A. Tantawi, "Performance modeling of distributed multi-tier enterprise systems,"in MAMA '09: Proceedings of the Performance Modeling of Distributed Multi-Tier Enterprise Systems, EleventhWorkshop on Mathematical Performance Modeling and Analysis, 2009.