A STRATEGIC PERSPECTIVE ON AUTONOMIC COMPUTING IN CLOUD

***Tarlochan Kaur, ** Ravneet Kaur**

** Assoc. Prof., Electrical Engineering Department*

PEC University of Technology, Chandigarh, India

***Software Engineer, Amadeus Labs, Bangeluru, India*

ABSTRACT

With the development of the internet, processing and storage technologies, computing resources have become cheaper, more powerful and available 24X7 everywhere. Cloud computing has now established itself as a preferred paradigm for hosting and delivering services over the internet. It allows consumers to start from the small and increase resources when there is a rise in service demand, eliminating the requirement to plan ahead for provisioning. One of the key challenges of cloud computing is the need for an automated and an integrated rational approach for dynamic provisioning of resources. This will allow the cloud to offer services that are reliable, cost-efficient and secure. In this context, autonomic computing promises to be a new concept for increasing reliability, autonomy and performance by enabling systems to adapt to changing circumstances. The aim of this paper is to provide a better understanding of the type of system architecture required to support above told objectives in cloud computing.

Keywords—Autonomic; Cloud computing; Resource provisioning

INTRODUCTION

Although the term 'cloud computing' evolved from concepts dating back to the sixties, it has created much hype and interest in the last few years. This was mainly because of recognition of its massive potential to contribute to the technological advances particularly in distributed computing.

Today cloud has been widely adopted as a service. Cloud computing is a distributed computing paradigm that is massively scalable and provides on demand services that are dynamically configurable. Users can demand almost unlimited computing power from anywhere with an internet connection, eliminating their need of making a major capital investment. Cloud computing delivers software (applications), platform and infrastructure services in a pay-as-you-go model. These services are referred to as Infrastructure as a Service (IAAS), Software as a Service (SAAS) and Platform as a Service (PAAS). This paradigm has made tremendous impact on the Information Technology (IT) industry, where we are witnessing corporate giants such as Google, Amazon and Microsoft striving to provide robust, reliable and cost-efficient cloud platforms. They have deployed cloud data centres around the world to provide the above said services. However the management of complex and heterogeneous cloud infrastructure is not only crucial but also a

challenging and an error-prone task. Due to the inherent dynamism, complexity and heterogeneity of clouds, they constitute an interesting venue to explore the uses of features of autonomic computing. It requires co-optimization at the cloud layers, thus exhibiting autonomic properties [1].

The main aim of autonomic computing is to realize computer and software systems that can manage themselves with little or no human interaction. Autonomic systems have the inherent capability of self-configuration, self-optimization, self-healing, hence self-management, along with constant tuning of their performance parameters. This enables them to free the system administrators from the details of system operations and maintenance and to provide users with a machine that with a machine that runs at peak performance 24X7 [2].

Although cloud has been widely adopted by the industry, the research on it, is still at an early stage. One of the grand challenges is the development of autonomic cloud system architecture that will increase the reliability, autonomy and performance by enabling systems to adapt to changing circumstances. This paper provides a better understanding of the system architecture, autonomic resource provisioning and management techniques to support such objectives in cloud computing.

CLOUD COMPUTING

Cloud computing "refers to a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interation" according to the definition provided by The National Institute of Standards and Technology (NIST) [3]. This definition captures the real essence of this emerging paradigm. It is a massively scalable, dynamically configurable and delivering different levels of services (SAAS, PAAS, and IAAS) to the customer on an on-demand basis.

One of the key features of cloud computing that makes it so attractive to service providers and business owners alike, is its ability to allocate and de allocate resources in accordance to demand. An immediate question is how these resources maybe managed so that it is cost effective? If resource allocation in static, then it may result in immense wastage of resources. Even if the service provider is able to correctly forecast the peak load, and provide capacity, in advance, to handle such load without dynamism and elasticity there is wastage of resources during no-peak time as shown in Figure 1.

To meet the QoS objectives and SLA requirements, the service provider tends to allocate resources in such a way that the worst-case demands are met. However, this over-provisioning of resources results in extra maintenance costs including server cooling and administration [4].

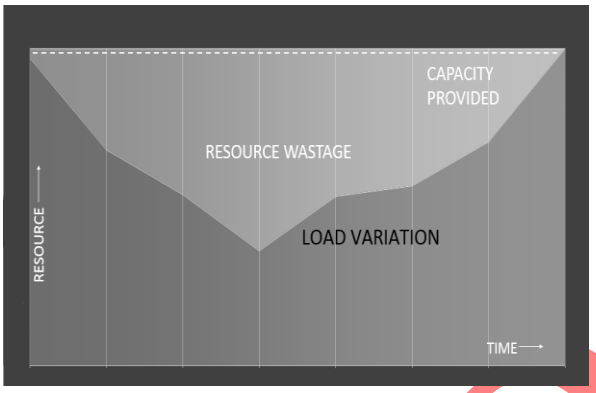

Fig. 1 Static resource allocation for peak loads

There may also be cases wherein the service provider underestimates the network load, and consequently, does not provide enough resources, turning away excess users. In the process he not only forgoes potential revenue from the users not served but also leaves the users unable to access the services as shown in Figure 2.

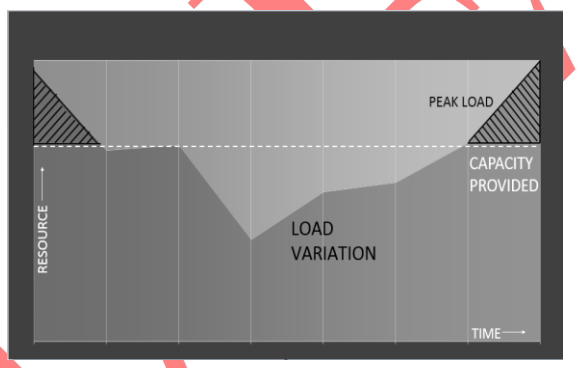

Fig. 2 Under provisioning of resources in static resource allocation

Due to under provisioning of resources, some users may desert the service provider permanently (after experiencing poor services) leading to permanent loss of load and a portion of revenue stream (Figure 3).

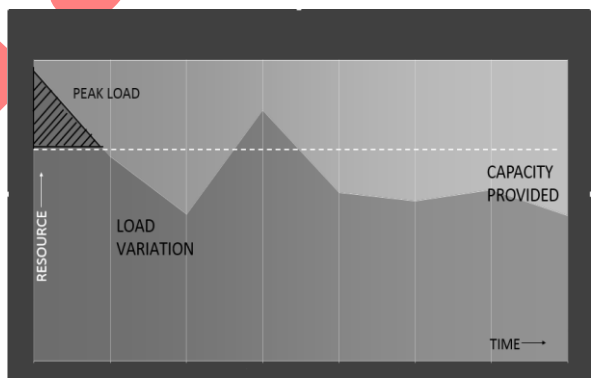

Fig. 3 Loss of traffic due to negative impact of under provisioning of resources in static resource allocation

In contrast, autonomic resource provisioning could lead to a faster response especially to a rapidly changing workload, and efficient resource utilization. With efficient application of its self-* properties, it can manage the complexity, and dynamism of cloud computing effectively.

AUTONOMIC COMPUTING

Autonomic Computing has become an essential aspect for a large scale distributed network. Even more so in cloud, as it has been built from the ground up to be a massively scalable, dynamic technology.

The objectives represent system requirements, while the attributes are representing tools for the implementation of the objectives [5]. Autonomic systems are self-configuring, self-healing, self-optimising and self-protecting. In other words, autonomic systems are self-managing. By virtue of self- healing property of the system, the system will detect and possibly repair, the localized software and hardware problems. It can achieve this through continuous monitoring, regression testing, using knowledge about system configuration etc. A self-protecting system will automatically defend itself against malicious attacks or cascading failures. It will anticipate and prevent system-wide failures based on early warnings and reports from sensors. Self-optimization is concerned with ensuring that the system runs at peak performance 24X7. This means it will proactively seek opportunities to upgrade its function by checking and applying the latest upgrades. Self-configuration refers to the system's abilities to configure itself automatically in accordance with the changing circumstances and the high-level objectives given by the system administrators. The high-level objectives, will simply specify what is desired and expected, and not how it is to be achieved. To achieve these objectives, the system has to be aware of both, its internal state, which can be achieved through self-analysis, as well as its external operating environment (environment-aware). For self-analysis, the system must have the ability to retrieve information on its internal state and behaviour. This is achieved through self-monitoring. Self-monitoring also facilitates detection of changing circumstances [6].When the change in the circumstances is detected, the system modifies its behaviour, configuration or even its internal structure (self-adjusting). To ensure that it is achieving its high-level goals, the system is continuously evaluating its performance against the ideal measure (self-critical).

ROLE OF AUTONOMIC COMPUTING IN CLOUD

There always exists the possibility that the user requirements can change overtime, which may thus lead to modifications in the original services requested by the user. The cloud infrastructure must be designed from the ground up in such a way that it can self-manage its limited resources (without compromising the QoS objectives and SLAs).

This can be facilitated through continuous supervision of services that are currently being requested, revising future service requests and incorporating changes in the schedules and prices for

the newly amended service requests. The system components must also be able to self-configure themselves in order to satisfy both, newly requested and revised service requirements. Thus an autonomic cloud architecture will be able to manage the resources in accordance to the stipulations of the service requirements.

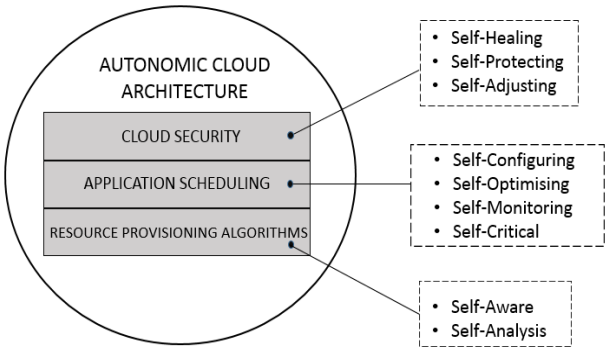

Fig. 4 Autonomic cloud architecture

Some aspects that have been identified for the implementation of autonomic resource provisioning are:

1. Developing scheduling mechanisms that will schedule cloud applications to cloud resources depending upon the QoS and application requirements. An application performance model will be constructed that will predict the number of cloud resources required to meet the demand. Also, the construction of the performance model will enable the mapping of the requirements to cloud resources ensuring effective allocation. Using the same model, prediction of future demand of the resources are made. The construction of the model is done autonomously by the cloud, using various techniques, such as Queuing Theory [7], Statistical Machine Learning [8], and Control Theory [9].

2. Implementation of dynamic provisioning algorithms: The algorithms has been studied extensively in the past [7, 10]. Using the performance model, and the predicted resource requirements, these algorithms automatically allocate cloud resources to the applications. This will not only ensure that the QoS objectives are being met but also the cost of services is being minimized for the user. Allowing resources to be managed autonomously, will help the service provider to optimize some objective function of interest to them (e.g. revenue) subject to some constraints as the workload varies in nature and intensity [11].The overall cost of using the services of the cloud would ultimately depend on the kind of resources that have been allocated along with the duration of their usage. The cloud will autonomously select the resources to be provisioned, by considering the past execution history and data of the cloud applications. The prediction would be based on the supply and demand for resources, similar to the market-oriented principles used for reaching equilibrium state [12].

3. Cloud security: The inherent complexity and heterogeneity of the cloud make it especially vulnerable to varies cyber-attacks, such as exploits, distributed denial of service attacks (DDoS) and so forth. In the context of resource provisioning, these attacks can cause unprecedented scaling-up of resources, by posing as authorized requests. This can lead to loss of revenue stream, inability to access cloud services etc. Hence the cloud must be able to autonomously distinguish legitimate from the illegitimate requests. By utilising, the self-healing properties, the system will protect itself against these malicious attacks. In case of suspicion, the cloud can either decide not to allocated resources altogether, or to avoid allocating excessive resources to the requests. Techniques that are already being used to handle such attacks can be moulded to fit into the characteristics of the cloud.

CONCLUSION AND FUTURE SCOPE

The growing popularity and demand of cloud computing in the case that it has established itself as the go-to solution for delivering and managing of services over the internet. This trend a testament of how important it has become to effectively manage the cloud resources. Albeit ground-breaking, the cloud technology is still in its early stages. Therefore there lies enormous potential of research and development in this field.

In this paper, a survey of how we can effectively manage the resources via autonomic computing has been made. The autonomic cloud platform presented in this paper will provide a better understanding of the type of system architecture required, thus paving way for more research in this area.

REFERENCES

[1] RajkumarBuyya, Rodrigo N Calheiros, and Xiaorong Li, "Autonomic cloud computing: open challenges and architectural elements," In Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on, pages 3-10. IEEE, 2012.

[2] Je_rey OKephart and David M Chess, "Thevision of autonomic computing," Computer, 36(1):41{50, 2003.

[3] Peter Mell and Timothy Grance. The Nist definition of cloud computing. NIST special publication, 800(145):7, 2011.

[4] Jong-Kook Kim, Howard Jay Siegel, Anthony A Maciejewski, andRudolf Eigenmann,"Dynamic resource management in energy constrained heterogeneous computing

systems using voltage scaling,"Parallel and Distributed Systems, IEEE transactions on, 19(11):1445{1457, 2008.

[5] Roy Sterritt and David W Bustard, "Autonomic computing - a means of achieving dependability," 2003.

[6] Roy Sterritt and David W Bustard, "Towards an autonomic computing environment," 2003.

[7] BhuvanUrgaonkar, Prashant Shenoy, Abhishek Chandra, and PawanGoyal, "Dynamic provisioning of multi-tier internet applications," In Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on, pages 217-228. IEEE, 2005.

[8] BhuvanUrgaonkar, Prashant Shenoy, Abhishek Chandra, and PawanGoyal, "Dynamic provisioning of multi-tier internet applications," In Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on, pages 217-228. IEEE, 2005.

[9] Evangelia Kalyvianaki, Themistoklis Charalambous, and Steven Hand, "Self-adaptive and self-configured cpu resource provisioning for virtualized servers using kalman filters."in proceedings of the 6th international conference on Autonomic computing, pages 117-126. ACM, 2009.

[10] Qi Zhang, LudmilaCherkasova, and EvgeniaSmirni,"A regression-based analytic model for dynamic resource provisioning of multi-tier applications in autonomic computing," 2007. ICAC'07, fourth international conference on, pages 27-27. IEEE, 2007.

[11] Emiliano Casalicchio, Daniel A Menasc_e, and ArwaAldhalaan, "autonomic resource provisioning in cloud systems with availability goals," in proceedings of the 2013 ACM Cloud and autonomic computing conference, page 1. ACM, 2013.

[12] RajkumarBuyya, Chee Shin Yeo, and SrikumarVenugopal,"Market-oriented cloud computing: vision, hype, and reality for delivering its services as computing utilities in high performance computing and communications," 2008. HPCC'08. 10th IEEE international conference on, pages 5-13. IEEE, 2008.