



Elastic Resource Allocation Challenges in Cloud computing System

¹Dr. M. Durairaj, ²P. Kannan

¹Assistant Professor, ²Research Scholar

School of Computer Science, Engineering and Applications,
Bharathidasan University, Tiruchirappalli, India.

Abstract: Elastic resource allocation in cloud computing paradigm is managing the resource elastically from anywhere and in anytime automatically. Elasticity is the fundamental characteristic by combining on-demand self service and resource pooling techniques in cloud, distributed, grid and utility computing circumstance. Realizing the significance of elasticity in cloud computing, this paper gives comprehensive analyses of elastic resource management and scheduling in cloud. In addition to, this paper discourse open challenges and future directions connected with role of elasticity mechanism in cloud computing.

Keywords: Cloud Computing, Elasticity, Resource Management, Partitioning, Scheduling, Virtualization.

I. Introduction

Cloud computing is extending the computing approaches from self model to service oriented access in which IT resources are consumed as on-demand and pay-per-usage model. Cloud computing elevates an on-demand elastic model for resource purveying, where resources can be an application, platform, hardware and virtual server utilized as services affirmed by data centers and cloud providers in a collaborative manner. Cloud based elastic resource provisioning [1] and management is a rapidly elaborating business, which reduces the delivery and data storage cost.

National Institute of Standards and Technology (NIST) has given a definition [2] for Cloud computing which says that “Cloud Computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources such as networks, servers, storage, applications and services that can be rapidly provisioned and released with minimal management effort or service provider interaction”. Five essential characteristics of cloud computing listed by NIST are on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service.

In recent years, cloud computing is explored and adopted by most of the companies, research institutions, academia and end users. By 2017, according to the CISCO global cloud index [3] forecasts nearly two-thirds of the proceeded transition of workloads will be processed from traditional data centers to cloud based data centers. The increments of workloads in cloud based data centers are awaited to be five-fold the growth in traditional workloads between 2012 and 2017.

The main reason to adopt cloud is the hypothesis of getting resources in an elastic manner. As a matter of fact, elasticity is a fundamental characteristic in the cloud and provides resources elastically on on-demand basis. The concept of cloud elasticity is the power to scale down and scale up of system resources based on workload variations of cloud users by adjusting the competence of cloud system resources.

Even though, cloud elasticity is frequently related with the scalability, an elusive difference presents between scalability and elasticity when observing a behavior of cloud system. Observing time intervals is one of the common ways to differentiate scalability and elasticity. Scalability is frequently used to depict the processes and mechanisms employed to a system workload with properties such as proper capacity planning, predictable approach and modifications which can be formed manually. In elasticity conception, where resources are utilized by instances and workload changes occur dynamically in a short period of time. In addition, resources are allocated flexibly with fully automated or minimum manual support.

According to NIST [4], the rapid elasticity is defined as “Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time”. Elasticity is done with the help of virtualization [5] which is a technique that creates the virtual machines by abstracting the system resources.

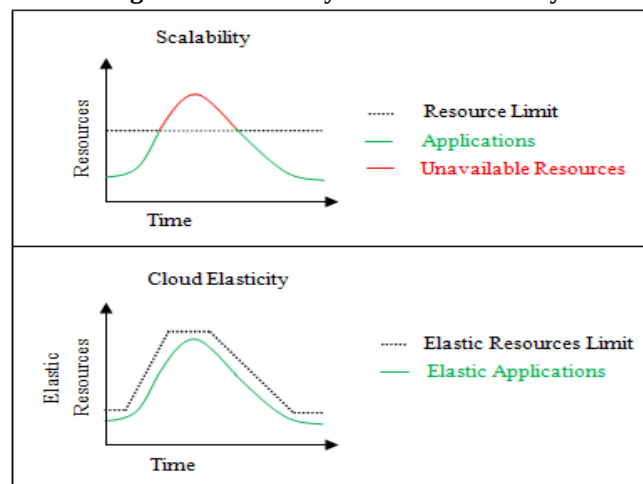
Elasticity should be broadly classified into two categories: application elastic nature and on-demand elastic resource utilization. First, the application elastic nature indicates static or dynamic characteristic of application and the ratio of the applications which are partitioned and offloaded to normal cloud or mobile cloud environment. Moreover, predictable approach will be applied to estimate the users limit for resource usage in normal and peak hours.

Second, the on-demand elastic resource utilization indicates that scaling of system resources or nodes are increases or decreases based on the availability and utilization of resources. Suitable scheduling techniques are applied for managing elastic resources to increase user access and reduce execution time and cost.

Figure 1 show that how cloud elasticity manages elastic resources and execution time in effective manner rather than traditional scalability approach. When resources reach their upper resource limit, the applications or infrastructures will not be able to access in traditional scalability systems. In cloud elasticity, resources are acquired from inter cloud providers and accessed without any constraints.

This paper focussed on how resources and applications are handled by elastic manner and suitable scheduling strategies which are available in cloud environment. Cloud elasticity provides resources in elastic manner with on-demand user access. From the perspective of cloud users, who pay cost on usage basis expects from service providers to support them to manage peak hour workloads without buying resources, increase cost savings, reduce manual support and maintain good rapport with his client. From the view of cloud providers who handle more clients simultaneously with needed resource requirement, increase revenue for their companies using cloud computing techniques and no need to buy resources frequently. Finally, we should focus on how elastic application and resources will be applicable for mobile cloud environment and three different types of cloud environment such as public, private and community cloud.

Figure 1: Scalability Vs Cloud Elasticity



The remaining of this paper is organised as follows: Section II presents works that are related to elastic resource management in cloud. Section III describes what kinds of scheduling challenges in cloud elasticity. Section IV concludes this paper and gives future direction of research.

II. Related works

At any point of time, cloud infrastructure resources are to be allocated to manage workload variations effectively and resources are shared among the cloud users in virtual mode with the help of elasticity. In order to achieve effective elastic resource management, it is necessitated for looking at the issues such as elastic resource provisioning, load balancing and resource scheduling and time management.

Chase, et al. [6] conceived the challenges towards energy oriented homogeneous resource management. The main challenge presented in this paper is to determining on-demand resources for user's requirement and provide resources in most convenient manner. Measuring of resource usages is very important in order to calculate elastic resource usage rating and provide best pricing model to cloud users and providers based on charge-back model which deals with system resource measurement and responsible for surplus capacitance needed for supporting an on-demand shared usage and dynamic model.

Rajkumar Buyya, et al. [7] presented that handling and monitoring of cloud resources energy consumption is based on elastic nature of cloud computing. Elasticity is one of the central facets in cloud computing which maintains good rapport with cloud users and providers. Zhu, et al. [8] depicted resource capacity and cloud workload scheme which combines different resource restrainers in automated resource allocation manner by applying three versatile time scales and scopes. Three kinds of time scales such as shortest time scale, shorter time scale and longest time scale are used. The controllers which are node controller, pod controller and pod set controller manage virtual machine resource allocation, workload migration, decides enough resource capacity for user requirements. Nevertheless, needed to concentrate on applying different threshold techniques for dynamic workload variations in cloud.

There are various exertion have used to scale elastic cloud application frameworks [9, 10] which alleviate elastic resource consumption using elastic services in cloud. This framework is developed to handle elastic services

based on autonomic computing precepts and proficiencies. However, there are no predictive techniques, algorithms to handle resource allocation and manage large quantities of data in elastic cloud services.

Bacigalupo, et al. [11] describes a resource scaling methods with the help of queuing models which contains three levels such as database, storage disk and application. Each levels are using queuing model result analysis which contains throughput, response time and resource utilization of a cloud server. By these analysis results, elastic cloud resource allocation algorithm is suggested to scale applications in dynamic environment. Marshall, et al. [12] presented elastic site which is a cluster component that supplies resources in elastic manner. It decides when to increase virtual machine count or decrease VMs count in cloud based on the data furnished by the elastic queue detector in order to manage the resource utilization effectively.

Elastic resource management scaling techniques [13] can be classified as predictive or reactive approach. Predictive approach means the elastic cloud system attempts to forecast succeeding resource demands in order to guarantee adequate cloud resources is uncommitted for on-demand purpose. It is based on heuristic proficiency, control theory, strengthener learning, time series analysis and analytical techniques. Using workload predictor and execution model to decide the number of cloud servers needed to handle the predicted system workload. A sort of predictive technique [14, 15, and 16] has been proposed for automatically scale cloud resources. Reactive approach indicates that the system responds to workload changes without pre planning and it focus scaling resource considerations is based on target measured time.

Elasticity handles the postulation for new cloud resources and relinquishes of idle resources. Typically, cloud environment have an elastic controller [17] which handles cloud user requisite conversion, monitoring applications data and decides the cloud resources must be scaled or not. But this controller is desirable only for client-server applications to manage workload.

Nowadays large-scale technological applications using the dynamically purveyed cloud resources as a form of Infrastructure as a Service (IaaS) instead of using company owned resources. For example, Science cloud [18] provides IaaS cloud resources to researchers and scientists who have no committed cloud computing resources. Amazon EC2 [19] provides virtual machine resources to execute variety of cloud users application on on-demand service basis from their infrastructure.

Genaud, et al. [20] examined various schemes based on existing heuristics to minimize resource utilization cost and waiting time of cloudlets or jobs for bin-packing problems and online scheduling in cloud. But they were not producing any new resource provisioning algorithms and scheduling policies for workload evaluation. Armstrong, et al. [21] applies minimum execution time to each cloudlet in arbitrary mode to the cloud nodes on which it is anticipated to be accomplished quickest, despite the current load on that node. Existing resource scheduling techniques such as round robin, first come first serve (FCFS) and min-min for load balance are used. Applying scheduling techniques [22, 23] in elastic clouds for better cost optimization to both cloud provider and user by conceiving service level agreement (SLA) violations and job priorities. Villegas, et al. [24] exhibited four resource allocation policies, eight resource provisioning techniques and cost analysis in a detailed study for IaaS clouds.

Computation partitioning methods and offloading techniques are major contributions to save energy usance for elastic resource management in mobile cloud computing environment. Generally, partitioning techniques are deals with a global and local view approach. In local view [25], the resource partition and offloading determination for each cloudlet or job is separately processed to reduce processing cost in cloud rather than traditional data transmission and without the support of global view.

In global view [26, 27 & 28] approach, resource computation partition and offloading mechanisms are interlinked with each cloudlet in the application. Partitioning and offloading challenges in mobile cloud environment are sculptured as an optimization problem. These problems intended to take one of the factors such as energy utilization cost, data transmission cost, scheduling criteria and latency to optimize applications.

Mobile cloud computation partitioning and offloading problem are taken into consideration from the view of single user in cloud to save energy utilization and improve application performances. The challenging new area of research is to examine how to improve elasticity in mobile cloud environment. This paper, presents the available challenges to schedule resources to produce effective and efficient elastic resource utilization in normal and mobile cloud.

III. Challenges

This segment discourses the challenges to improve resource utilization in normal and mobile cloud computing systems. Many of the disputes that are confronted in elastic virtual resource management using cloud resource allocation techniques, scheduling strategies and load balancing in order to provide virtual machines distribution among servers or datacenters to attain the most beneficial of cloud system performance.

Elastic Load balancing: Cloud providers are normally using load balancing services, which permit cloud resources or load to be increased or decreased based on the demands without considering the elasticity. The resources can be storage, CPU, network bandwidth limit, memory and applications. Elasticity is one of the prerequisite for enhancing the cloud computing operation and utilizes resources in elastic manner. Load balancing of virtual machine resources is a challenging task in cloud environment due to the dynamic nature.

For providing a better load balancing the concentration should be on qualitative metrics such as elasticity, throughput, fault tolerant, overhead, resource utilization, response time, migration time and performance in normal and mobile cloud environment using load balancing algorithms. Existing algorithms are not concentrating on elasticity metrics in load balancing for better resource utilization. Comparative studies of existing load balancing algorithms depicted in table 1.

Token routing algorithm is focusing the movement of tokens in cloud system but not having adequate information to distribute the workload in elasticity cloud environment. This algorithm furnishes the quick and better routing decision. Round robin is one of the mere resource load balancing techniques that applies the rationale of time slices which handles cloud users request based on particular time interval or slice. Throttled load balancer algorithm allocates determined cloudlets, which contains job or task depends on the environment to a single virtual machine. Moreover, it is using queue mechanism when more requests occur at a time.

Genetic algorithm is applied to reduce virtual machine scheduling time and improve response time using population, crossover and mutation techniques but the migration time becomes a problem. Randomized algorithm handles the process based on the probability value. This algorithm works well when the resources are equally distributed and not exerting deterministic procedure. Central Queuing algorithm exploits on the concepts of dynamic workload distribution and having activity, queue controller, requester and load manager.

Table 1: Comparison of Major Scheduling Algorithms with Metrics

Algorithms Metrics	Token Routing	Round Robin	Throttled Load Balancer	Genetic	Randomized	Central Queuing	Priority Based Scheduling	Gang Scheduling	Greedy Algorithm
Dynamic Nature	✓	x	✓	✓	x	✓	x	✓	x
Stability	x	✓	x	✓	✓	x	x	x	x
Migration	✓	x	x	x	x	x	✓	✓	x
Response Time	✓	✓	✓	✓	✓	✓	✓	✓	✓
Scalability	x	x	x	✓	x	x	x	✓	✓
Trust	x	x	x	✓	x	x	x	x	x
Execution Time	✓	✓	✓	✓	✓	✓	✓	✓	✓
SLA	x	✓	x	x	x	x	✓	x	x
Resource Utilization	✓	✓	✓	✓	x	x	x	x	x
Elastic Load Balancing	x	x	x	x	x	x	x	x	x
Fault Tolerant	x	x	x	x	x	x	x	✓	x
Cost	✓	✓	✓	✓	x	✓	✓	✓	x

Gang scheduling algorithm is accountable for choosing of finest desirable resources for VM or cloudlet execution for scheduling parallel jobs by applying static and dynamic metrics in cloud environment. Priority based scheduling algorithm allocates virtual machines based on workload weight, cloud user allotted priority, resource utilization time, migration and execution time. Priority altered dynamically to virtual machines by aging techniques.

Elastic Resource Scheduling is the canonical and central part for managing and providing cloud resources effectively and efficiently according to user demands as well as cloud service provider's capacity. Based on Service Level Agreements, VM or jobs requires scheduling and allocating cloud resources by improving Quality of Service. None of the existing scheduling and load balancing algorithms concentrates on elasticity metrics to maximize resource utilization, consume adequate energy, reduce virtual machine migration count, time and preserve cost to cloud users and providers.

IV. Conclusion and Future Work

This review paper presents the versatile works on scheduling and resource management algorithms in cloud computing technology to improve elastic load balancing and reduce virtual machine migration time or count with its challenges. The existing scheduling algorithms and metrics that have not been applied to integrate a cloud elastic framework for load balancing, resource allocation and scheduling strategy in normal cloud and mobile cloud environment are listed in Table 1, which is a major objective of this work. The proposed framework considers the cloud users and providers constraints such as elastic load balancing, migration time, dynamic nature, response time and cost. This research work disclosed that none of the existing scheduling strategies surveyed could accomplish the elastic resource management metrics. As a future work, we propose a framework to manage elasticity in application and infrastructure level to improve resource utilization, load balancing and virtual machine migration time in normal and mobile cloud environment effectively.

References

- [1]. M. Durairaj, P. Kannan, "A Novel Approach for Elastic Application Partitioning in Mobile Cloud", IEEE - ICAET-4th International Conference on Advances In Engineering & Technology, India, 2014.

- [2]. P. Mell, T. Grance, "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, Information Technology Laboratory, Technical Report Version 15, 2009.
- [3]. White Paper, "Cisco Global Cloud Index: Forecast and Methodology 2012–2017", CISCO, 2013.
- [4]. P. Mell, T. Grance, "The NIST Definition of Cloud Computing", Tech. rep., U.S. National Institute of Standards and Technology (NIST), 2011. Special Publication 800-145, <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.
- [5]. M. Durairaj, P. Kannan, "A study on Virtualization Techniques and Challenges in Cloud Computing", International Journal of Scientific & Technology Research, Volume 1, Issue 1, 2014.
- [6]. JS Chase, Darrell C Anderson, Prachi N Thakar, Amin M Vahdat, "Managing energy and server resources in hosting centers", In: Proceedings of 11th IEEE/ACM international conference on grid computing (GRID), vol.12, no.4, 2010.
- [7]. R. Buyya, R. Ranjan, and R. Calheiros, "Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services." In ICA3PP 2010: The 10th International Conference on Algorithms and Architectures for Parallel Processing, 2010, pp. 19–24.
- [8]. Xiaoyun Zhu, Donald Young, Brian J. Watson, Zhikui Wang, Jerry Rolia, Sharad Singhal, Bret McKee, Chris Hyser, Daniel Gmach, Robert Gardner, Tom Christian, Ludmila Cherkasova, "1000 islands: Integrated capacity and workload management for the next generation data center", Proceedings of the 5th International Conference on Autonomic Computing (ICAC 2008), Springer Science, Chicago, IL, USA, 2008.
- [9]. P. Martin, A. Brown, W. Powley, J.L. Vazquez-Poletti, "Autonomic management of elastic services in the cloud", in: IEEE Symposium on Computers and Communications, ISCC'11, IEEE, Kerkyra, 2011.
- [10]. E. Casalicchio, L. Silvestri, "Architectures for autonomic service management in cloud-based systems", in: IEEE Symposium on Computers and Communications, ISCC'11, IEEE, Kerkyra, 2011.
- [11]. D.A. Bacigalupo, J. van Hemert, X. Chen, A. Usmani, A.P. Chester, D.N. Dillenberger, G.B. Wills, L. Gilbert, S.A. Jarvis, "Managing dynamic enterprise and urgent workloads on clouds using layered queuing and historical performance models", Simulation Modelling Practice and Theory, 2011.
- [12]. P. Marshall, K. Keahey, T. Freeman, Elastic site: using clouds to elastically extend site resources, in: Proceedings of the 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2010.
- [13]. T. Llorido-Botrán, J. Miguel-Alonso and J.A. Lozano, "Auto scaling Techniques for Elastic Applications in Cloud Environments," Technical Report EHU-KAT-IK-09-12, 2012.
- [14]. W. Dawoud, I. Takouna, and C. Meinel, "Elastic vm for cloud resources provisioning optimization," in Advances in Computing and Communications, ser. Communications in Computer and Information Science, A. Abraham, J. Lloret Mauri, J. F. Buford, J. Suzuki, and S. M. Thampi, Eds. Springer Berlin Heidelberg, 2011, vol. 190, pp. 431–445.
- [15]. N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in Proceedings of the 4th Intl. Conference on Cloud Computing, ser. CLOUD 2011. IEEE, 2011, pp. 500–507.
- [16]. U. Sharma, P. Shenoy, S. Sahu, and A. Shaikh, "A cost-aware elasticity provisioning system for the cloud," in Proceedings of the 31st Intl. Conference on Distributed Computing Systems, ser. ICDCS 2011. IEEE, 2011, pp. 559–570.
- [17]. L. M. Vaquero, L. Roderio-Merino, and R. Buyya, "Dynamically scaling applications in the cloud," SIGCOMM Comput. Commun. Rev., vol. 41, pp. 45–52, 2011.
- [18]. Sciencecloud, <https://www.sciencecloud.com>
- [19]. Amazon: Amazon Elastic Compute Cloud, <http://aws.amazon.com/ec2/>
- [20]. S. Genaud, J. Gossa, "Cost-Wait Trade-Offs in Client-Side Resource Provisioning with Elastic Clouds", Cloud Computing (CLOUD), 2011 IEEE International Conference on, July 4–9, 2011, Washington DC, USA, IEEE, 2011:1–8.
- [21]. T. R. Armstrong, D. Hensgen, "The relative performance of various mapping algorithms is independent of sizable variances in runtime predictions", in Proc. of 7th IEEE Heterogeneous Computing Workshop (HCW 98), pp. 79–87, 1998.
- [22]. M. Hyun Jin, C. Yun, et al. SLA-Aware Profit Optimization in Cloud Services via Resource Scheduling. Services (SERVICES-1), 2010 6th World Congress on. July 5–10, 2010, Miami, USA, 2010: 152–153.
- [23]. P. Leitner, W. Hummer, et al. Cost-Efficient and Application SLA Aware Client Side Request Scheduling in an Infrastructure-as-a-Service Cloud. Cloud Computing (CLOUD), 2012 IEEE 5th International Conference on, June 24–29, 2012, Honolulu, Hawaii, USA: 213–220.
- [24]. D. Villegas, A. Antoniou, et al. An Analysis of Provisioning and Allocation Policies for Infrastructure-as-a-Service Clouds. Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on, May 13–16, 2012 : 612–619.
- [25]. Z. Li, C. Wang, and R. Xu. Computation offloading to save energy on handheld devices: a partition scheme. In Proceedings of the International Conference on Compilers, Architecture, and Synthesis for Embedded Systems, pages 1–6. IEEE Press, 2001.
- [26]. L. Yang, J. Cao, S. Tang, T. Li, and A. Chan. A Framework for Partitioning and Execution of Data Stream Applications in Mobile Cloud Computing. In Proc. of IEEE Fifth International Conference on Cloud Computing. 2012.
- [27]. X. Zhang, A. Kunjithapatham, S. Jeong, S. Gibbs. Towards an elastic application model for augmenting the computing capabilities of mobile devices with cloud computing. Mobile Networks and Applications, Volume 16, Issue3, 2009.
- [28]. E. Cuervoy, A. Balasubramanian, and D. Cho. MAUI: Making Smartphones Last Longer with Code Offload. In Proc. of MobiSys'10. ACM press, 2010.