Asian Journal of Science and Applied Technology
ISSN: 2249-0698 (P) Vol.1 No.2, 2012, pp.31-35
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajsat-2012.1.2.732

Survey on Resources Provisioning in Cloud Systems for Cost Benefits

M. Karthi and S. Nachiyappan
Department of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India
E-mail : karthi.vcet11@gmail.com, esspee27@gmail.com

(Received on 14 August 2012 and accepted on 15 October 2012)

Abstract – **Cloud providers can offer cloud consumers two provisioning plans for computing resources, namely reservation plan and on-demand plan. In generally, the cost of utilizing computing resources provisioned by reservation plan is cheaper than on demand plan. There are many kinds of resource provisioning options available in cloud environment to reduce the total paying cost and better utilizing cloud resources. However, the best advance reservation of resources is difficult to be achieved due to uncertainty of consumer's future demand and providers' resource prices. To address this problem Probabilistic based cloud resource provisioning (PCRP) algorithm is proposed by formulating a Probabilistic model. In this paper survey the different provisioning options andalgorithm. Compare the existing provisioning algorithms with analysis based on cost, availability, uncertainty parameters.**

Keywords: **Cloud Computing, Future Demand, Integer Programming, Resource Management, Resource Provisioning**

I. Introduction

Cloud computing is the utilization of computing software and hardware resources. Those Resources are delivered to cloud consumer as a service over a network typically the Internet. There are three major categories of cloud services is Infrastructure as a service (IaaS),Platform as a service (PaaS) and Software as a service (SaaS). The most basic cloud-service model, providers of IaaS offer computers-physical or virtual machines and some other resources (images in a virtual-machine image-library, file-based storage, raw (block) and firewalls, IP addresses, load balancers, virtual local area networks, and some of the software bundles). In the PaaS model, cloud providers deliver a computing platform typically including programming language execution environment, database, operating system, and web server. In the SaaS model cloud providers operate application software in the cloud and cloud users access the software from cloud clients. The cloud users normally do not manage the cloud infrastructure and platform on which the application is running. Toeliminate need to install and run the application on the cloud user's own computers simplifying maintenance and support.

In cloud computing a resource provisioning mechanism is required to supply cloud consumers a set of computing resources for processing the jobs and storing the data and etc. Cloud providers can offers to cloudconsumer's two resource provisioning plans. That namely short-term on-demand and long-term reservation plans. Amazon EC2 and Go Grid arecloud providers which offer IaaS services with both plans. In generally pricing in on-demand plan is charged by pay-per-use basis (e.g., per day, per hour). Therefore purchasing this on-Demand plan, the consumers can dynamically provisioning resources at the moment when the resources are needed to fit the fluctuated and unpredictable and unexpected demands. For reservation plan, pricing is charged by a onetime fee (e.g., 1 year, 3year) typically before the computing resource will be utilized by cloud consumer. In the reservation plan the price to utilize resources is cheaper than that of the on-demand plan. So the consumer can reduce the cost of computing resource provisioning by using the reservation plan. In the reservation plan offered by Amazon EC2 can reduce the total provisioning cost up to 50 percent approximately when the reserved resource is fully utilized at steady-state usage. With the reservation plan, the cloud consumers a priory reserve the resources in advance. As a result the under provisioning problem can occur when the reserved resources are unable to fully meet the demand due to its uncertainty. This problem can be solved by provisioning more resources with on-demand plan to fit the additional demand, the high cost will be incurred due to more expensive price of resource provisioning with on-demand plan. On the other hand the over provisioning problem can occur if the reserved resources are more than the actual demand in which part of a resource pool will be underutilized. This is an important for the cloud consumer to minimize the total cost of resource provisioning by reducing the on-demand cost and oversubscribed cost of under provisioning and over provisioning. To achieve this goal, the optimal computing resource management is the critical issue.

II. Literature Survey

A. Resource Provisioning Options For Large-Scale Scientific Workflows

Advance reservations enable users to allocate resources for their exclusive use for a given period of time [1]. This technique reduces queuing delays by eliminating competition for shared resources To avoid or minimize job delays, several resource provisioning options are available to workflow applications Multi-level scheduling is a provisioning technique that enables user-level resource managers to control jobs and resources. This approach reduces queuing delays by reserving resources, and reduces scheduling delays by allowing scheduling policies to be managed at the application level. It does not provide solution for complete resource provisioning problem and also this algorithm does not concentrate on security management technique when resource provides to the users on the cloud.

B. Autonomic Provisioning Of Backend Databases In Dynamic Content Web Servers

In autonomic provisioning a resource manager will allocate resources to an application, on-demand, e.g., during load spikes[3]. Modeling-based approaches have proved very successful for provisioning the web and application server tiers in dynamic content servers. On the other hand, accurately modeling the behavior of the back-end database server tier is a daunting task .novel pro-active scheme based on the classic K-nearest-neighbors (KNN) machine learning approach for adding database replicas to application allocations in dynamic content web server clusters. KNN algorithm uses lightweight monitoring of essential system and application metrics in order to decide how many databases it should allocate to a given workload. KNN used to improve prediction accuracy and avoid system oscillations. K-nearest-neighbors (KNN) machine learning approach only, so this approaches not suitable for all clients in the cloud.

C. Variety of Science Applications are Integrated With Large-Scale HPDC

HPDC resource management paradigm named resource slot which defines a network of logical machines across space and time. A resource slot is not only a resource programming target but also a virtualized resource provisioning framework for a variety of resource management paradigms by encapsulating the resource management complexity. A resource provisioning technique was guided redundant submission (GRS).It probabilistically guarantees a timely resource slot allocation.Guided redundant submission which probabilistically secures the temporal resource availability, based on a joint failure probability of individual resources. This paper has not handle techniques for reduce the cost of resource.This paper follows the technique guided redundant Submission (GRS) therefore we cannot give complete surety for resource provides to the user in grid.

D. Risk-Aware Limited Look Aheadcontrols for Dynamic Resource Provisioning

In the resource provisioning problem is posed as one of sequential decision making under uncertainty and solved using a limited look ahead control scheme. The proposed approach used for the switching costs incurred during resource provisioning and explicitly encode risk in the optimization problem. The LLC concept is adopted from model predictive control, sometimes used to solve optimal control problems for which classical feedback solutions are extremely hard or impossible to obtain. The LLC approach is a practical option for enforcing self-managing behavior in resource provisioning applications. This paper follows dynamic approach for resource provisioning therefore small users may wait long time for get resource. No functionality for reduce the operating cost and operating time.

E. Sla-Aware Virtual Resource Management for Cloud Infrastructures

Ability to automate the dynamic provisioning and placement of VMs taking into account both application-level SLAs and resource exploitation costs with high-level handles for the administrator to specify trade-offs between the two. Support for heterogeneous applications and workloads including both enterprise online applications with stringent QoS requirements and batch-oriented CPU intensive applications. Support for arbitrary application topology: single cluster, n-tier, monolithic and capacity to scale: either in a"scale-up" fashion by adding more resource to a single server or in a "scale-out" fashion by adding more servers. It is not focused on optimization problems that are NP-hard in their general form.

F. A Hybrid Particle Swarm Optimization Algorithm for Optimal Task Assignment in Distributed Systems

In a computation system with a number of distributed processors, it is desired to assign application tasks to these processors such that the resource demand of each task is satisfied and the system throughput is increased. In a distributed system, a number of application tasks may need to be assigned to different processors such that the system cost is minimized and the constraints with limited resource are satisfied. The assignment of tasks will also incur some costs such as the execution cost and the communication cost. The task assignment problem (TAP) is to find an assignment of tasks which minimizes the incurred costs subject to the resource constraint. Most of the existing formulations for this problem have found to NP-complete and thus finding the exact solutions is computationally intractable for major large-scaled problems. The Hybrid particle swarm optimization algorithm used for finding the nearest optimal task assignment with reasonable less time. The experimental results manifest that the proposed method is more effective and efficient than a genetic algorithm. Our method converges at a fast rate and is suited to large-scaled task assignment problem.

G. Optimal Virtual Machine Placement across Multiple Cloud Providers

The under provisioning problem can occur when the reserved resources are unable to fully meet the demand due to its uncertainty. The over provisioning problem can occur if the reserved resources are more than the actual demand in which part of a resource pool will be underutilized . To solve this problem Optimal Virtual Machine Placement (OVMP) algorithm is optimally allocating VMs to multiple cloud providers and follows optimally advance reservation. In OVMP consider uncertainty of demand and price. This algorithm achieved by stochastic integer programming with two stage resources. Multiple VM class used each VM class have a distinct type of resources. Each virtual machine in VM class has different resource requirement. The number of VMs in each VM class depends on the demand from user. In first stage reserve the cheaper resources by using advance resource provisioning.in second stage consist of two phase utilization phase and on demand phase. Pay cost more to additional resources in on demand phase. Formulating the integer programming and solve it get optimal solution.

H. Optimization of Resource Provisioning Cost in Cloud Computing

The advance reservation of resources is difficult to be achieved optimized cost due to uncertainty of consumer's future demand and providers resource prices. The under provisioning problem can occur when the reserved Resources are unable to fully meet the demand due to its uncertainty. The over provisioningproblem can occur if the reserved resources are more than the actual demand in which part of a resource pool will be underutilized.To address this problem we use an optimal cloud resource provisioning (OCRP) algorithm is proposed by formulating a stochastic programming model. The OCRP algorithms can provisioningcompute resources for being used in multiple provisioning stages as well as in long-term plan, e.g., three stages in a quarter plan and twelve stages in a yearly plan. Different approaches used to obtain the solution of the OCRP algorithm are considered including sample-average approximation, deterministic equivalent formulation and Benders decomposition. Numerical studies are performed in which the results clearly show that with the OCRP algorithm, cloud consumer can able to successfully minimize total cost of resource provisioning in cloud computing environment.

III. Comparison of Provisioning Algorithm

The comparison between provisioning algorithms is performed as follow. The algorithms include the OCRP, expected-value of uncertainty provisioning (EVU), no reservation provisioning (NoRes) and maximum advance reservation provisioning (MaxRes) algorithms. EVU uses the average values of uncertainty parameters in cost and solves them by a traditional deterministic program. In MaxRes reserves the maximum number of available VMs to user he cloud resources. Both MaxRes and NoRes also apply the traditional deterministic program for allocating VMs to cloud providers.WhileNoRes does not reserve any resources to user for utilizing the probabilistic distributions are applied to the possible scenarios of price and demand respectively. The solution obtained from each solved algorithm yields the number of reserved VMs and the allocation of VMs to providers. The provisioning costs incurred by purchasing the provisioning plans given by the solution of each algorithm are recorded. Take sample input the costs include reservation cost (R.C.), expending cost (E.C.), on-demand cost (O.C.), oversubscribed cost (OS.C.), and total cost.

The OCRP achieves the lowest optimal total cost, while NoRes yields the highest total cost due to the highest on-demand resource cost. The OCRP algorithm reserves 60 VMs (including both classes I1 and I2). Although MaxRes reserves 100 VMs (50 per VM class) to entirely have higher cost in the on-demand plan, and incurs much higher cost than that of OCRP. Additionally MaxRes incurs the highest overprovisioning cost since the reservedresources are unnecessarily overprovisioned. EVU incurs the total cost lower than those of MaxRes due to take average number of VMs and NoRes algorithms. Although theoverprovisioning cost of OCRP is higher than that of EVU but the on-demand cost of the OCRPalgorithm is much lower than other algorithms. Again, it is possible that the on demand cost can increase due to the price uncertainty. As a result theuse of the on-demand cost is more important. The result of this comparison shows the balance between the number of provisioning resources to be acquired in the first and next stages in which OCRP can provide the most optimal cost.

IV. Conclusion

We suggest a probabilistic resource provisioning approach that can be exploited as the input of a dynamic resource management scheme. Using resource on Demand use case to justify our claims and to represent sudden workload variations we propose an analytical model inspired from standard models. Using this model we can eliminate under provisioning and overprovisioning problems.

References

[1] G. Juve and E. Deelman, “Resource provisioning options for large-scale scientific workflows,” *Proc. IEEE fourth int’l conf. e-science,* 2008.

[2] J. Chen, G.Soundararajan, and C.Amza, “Autonomic Provisioning of Backend Databases in Dynamic Content Web Servers,” *Proc. IEEE int’l conf. autonomic computing,* 2006.

[3] Y.Kee and C.Kesselman, “Grid resource abstraction, virtualization, and provisioning for time-target applications,” *proc. IEEE int’l symp. Cluster computing and the grid,* 2008.

[4] *D. Kusic and N. Kandasamy*, “Risk-aware limited look ahead control for dynamic resource provisioning in enterprise computing systems,” *Proc. IEEE int’l conf. autonomic computing,* 2006.

[5] H.N. Van, F.D. Tran, and J.M. Menaud, “Sla-aware virtual resource management for cloud infrastructures,” *Proc. IEEE ninth int’l conf. computer and information technology,* 2009.

[6] Peng-yeng yin, Shiuh-sheng yu, Pei-peiwang, “A hybrid particle swarm optimization algorithm for optimal task assignment in distributed systems”, *Journal computer standards & interfaces archive,* Vol. 28, No. 4, April, 2006, pp. 441-450.

[7] S.Chaisiri, B.S.Lee, and D.Niyato, “Optimal virtual machine placement across multiple cloud providers,” *Proc. IEEE asia- pacific services computing conf.* (APSCC), 2009.

[8] S.Chaisiri, B.S.Lee, and d. Niyato, "Optimization of resource provisioning cost in cloud computing,", *IEEE transactions on services computing,* Vol. 5, No. 2, April-June 2012.

[9] Y. Jie, Q. Jie, and L. Ying, "A profile-based approach to just-in-time scalability for cloud applications," *Proc. IEEE int'l conf. cloud computing (*CLOUD '09), 2009.

[10] K.Miyashita, K.Masuda, and F. Higashitani, "Coordinating service allocation through flexible reservation," *IEEE trans. services computing,* Vol. 1, No. 2, pp. 117-128, Apr.-June 2008.

[11] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing sla violations," *Proc. IFIP/IEEE int'l symp. Integrated network management* (im '07), pp. 119-128, May 2007.

[12] Wenjun Wu ; Dichen Di ; Fei Zhang ; Yizhou Yan ; Yaokuan Mao, "A resource scheduling algorithm of cloud computing based on energy efficient optimization methods", *Green Computing Conference* (IGCC), 2012, June 2012.

[13] Pradeep.R, Kavinya.R, "Resource Scheduling In Cloud Using Bee Algorithm for Heterogeneous Environment IOSR", *Journal of Computer Engineering (IOSRJCE) (July-Aug. 2012).*

[14] Yong Beom Ma, Sung Ho Jang, Jong SikLee, "Ontology-Based Resource Management for Cloud Computing", *Intelligent Information and Database Systems,* Vol. 65, 2011, pp. 343-352.

[15] Preeti Agrawal and Yogesh Rathore, "An Approach for Effective Resource Management in Cloud Computing," *Int. J. Tech. 2011,* Vol. 1, No. 2, pp. 121-124.

[16] Bahman Javadi, Parimala Thulasiraman and Rajkumar Buyya, "Cloud Resource Provisioning to Extend the Capacity of Local Resources in the Presence of Failures," *IEEE 14th International Conference on High Performance Computing and Communications,* 2012.

[17] Thomas sandholm, "Evaluating demand prediction techniques for computational markets", *Proceedings in GECON2006.*