A Survey on Resource Provisioning Schemes in Cloud

Chethan Venkatesh[1], Shiva Murthy.G[2]
[1]*Department of Computer Application, M.S.Ramaiah Institute of Technology Bangalore*
[2] *Department of MCA, VTU Center for PG studies, Bangalore*
chethu4u@gmail.com[1] *kgshivam@gmail.com*[2]

Abstract

Cloud computing has emerged as a major area in the field of computer and information technology. The area serves as major benefits for users as it avoids huge expenditure on hardware and software resources from the end user point of view. It adopts pay as you use model and the resources can be used as per the requirement and the end user will be billed accordingly. With all these advantages there are certain shortcomings with this model related to the provisioning of resources in an optimized manner with respect to cost and balancing the load. In addition we also need to look at the provisioning schemes in the event of failures. In this paper we present a survey which would give an insight of different provisioning schemes based on cost optimization, fault tolerance and load balancing issues.

Key words – Resource provisioning, Cost optimization, Fault Tolerance, load balancing.

1. Introduction

The word "cloud" often refers to the Internet. A Cloud is a type of distributed and parallel system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned. Cloud is presented as one or more unified computing resources based on service level agreements (SLA) established through negotiation between the service provider and consumers.

Resource provisioning is the allocation of a cloud provider's resources to a customer. When a cloud provider accepts a request from a customer, cloud provider must create the appropriate number of virtual machines (VMs) and allocate resources to support each one of them.
Resource Allocation Strategy (RAS) is all about integrating cloud provider activities for utilizing and allocating scarce resources within the limit of cloud environment so as to meet the needs of the cloud application.

The biggest benefit of resource allocation is that user neither has to install software nor hardware to access, develop, and host the application over the internet Irrespective of the location. We can reach our applications and data anywhere in the world, on any system without the need to expend on hardware and software systems. Cloud providers can share their resources over the internet during resource scarcity[1][16].
Since users rent resources from remote servers for their purpose, they don't have control over their resources. When the users wants to switch to some other provider for the better storage of their data. It's not easy to transfer huge data from one provider to the other. In public cloud, the clients' data can be susceptible to attacks like hacking or phishing. Since the servers on cloud are interconnected, it is easy for malware to spread. Peripheral devices like printers or scanners might not work with cloud as most of them require software to be installed locally. Networked peripherals have comparatively lesser problems [1].

In depth knowledge is required for provisioning of resources in cloud, since all knowledge about the working of the cloud mainly depends upon the cloud service provider.

In this paper we will discuss about the various resource provisioning schemes in cloud proposed by several researchers which will help us identifying key challenges in provisioning resources in an cloud environment.

2. Key Challenges In Resource Provisioning:

Scalability: As more and more end systems, other network devices join and leave the network, an efficient scalable resource allocation strategy is required.

Load Balancing: when there are multiple instances of resources across the cloud, resource requests must be distributed across all instances uniformly (virtual machine migration to balance load across the data center [9]).

Security: protection of various resources against destructions and unauthorized access.

IJERT

Fault-Tolerance: System should continue to function in the event of failures by allocating alternate resources / auto - recovery.

Cost Optimization: Resource provisioning is done in a cost effective manner, i.e., at minimal cost, by allocating the resource from shortest possible node available.

Bandwidth Issues: bandwidth reservation and provisioning is required for QoS (Quality of Service) in cloud computing.

Automated Service provisioning: The objective of a service provider in this case is to allocate and de-allocate resources from the cloud to satisfy its service level objectives (SLOs), while minimizing its operational cost [9].

Server Consolidation: Server consolidation is an effective approach to maximize resource utilization while minimizing energy consumption in cloud computing environment [9].

Energy management: Improving energy efficiency is another major issue in cloud computing. It has been estimated that the cost of powering and cooling accounts for 53% of the total operational expenditure of data centers [9].

Traffic Management and analysis: Analysis of data traffic is important for today's data centers. For example, many web applications rely on analysis of traffic data to optimize customer experiences. Network operators also need to know how traffic flows through the network in order to make many of the management and planning decisions [9].

3. Parameters for Resource Provisioning and its importance:

Several researchers are working in this area; a lot of work is being carried out with respect to several parameters for resource provisioning. In cloud computing, Resource Allocation (RA) is the process of assigning available resources to the needed cloud applications over the internet. Resource allocation starves services if the allocation is not managed precisely [1].

We have described various models proposed by several researchers and analysed its advantages and limitations below.

- Xiaoqiao Meng et al[8] have proposed a approach of virtual machine (VM) multiplexing and joint-VM provisioning.

  This approach considers SLA imposed on VM capacity, an algorithm which calculates total capacity needs for multiplexed VMs and to identify compatible VM combinations for consolidation and joint provisioning. The proposed method and its applications are evaluated by performance data collected from about 16 thousand VMs in commercial data centers. The results demonstrate more than 45% improvements in terms of the overall resource utilization.

- The work proposed makes an attempt to distribute requests fairly but does not address network failures and since it increases resource utilization, the cost factor may have to be compromised. Sivandon Chaisiri et al.[6] have proposed an Optimal Cloud Resource provisioning (OCRP) algorithm by formulating a stochastic programming model. It considers demand and price uncertainty as parameters and three phases of provisioning namely reservation, expending and on-demand. The algorithm aims at minimizing the total cost of resource provisioning for a cloud consumer.

  - As the cost factor is an important parameter in resource provisioning the OCRP algorithm can be of great importance. Optimization can be done by provisioning the resources from the nearest neighbor which may not guarantee fairness of allocation resulting in low resource utilization.

- M.Dakshayini et al.[15] have proposed Adaptive Fault Tolerant Service Provisioning [AFTSP] Strategy to achieve high system performance coupled with high fault tolerance and guaranteed QoS.

  - The cost optimization is very much necessary when it comes to fault tolerant systems, in case of a failure the system should recover in quick time by choosing alternate resource instance. The alternate solution is to be chosen in an optimized manner.

- Bahman Javadi et al.[11]have proposed a generic resource provisioning model based on the stochastic analysis of routing in distributed parallel queues where the arrival and service processes follows general distribution. The proposed model considers cost factor, failure aware scenario and scheduling.

IJERT

- The proposed model above considers the parameters in case of failure of local cluster. The model does not emphasize more on the load balancing issue which happens to be another important parameter. We can also extend the fault tolerant issue outside the local cluster.

- Bhaman Javadi et al.[12] have also proposed failure-aware resource provisioning for hybrid cloud infrastructure which address the QoS parameters of the end user. It considers the workload model, but the cost parameter is not considered for the model proposed.

  - The proposed model Considers the workload model, but the cost parameter is not considered which can be a limitation.

Several research issues are addressed with respect to resource provisioning in cloud (private, public, hybrid, heterogeneous), prominent among them are cost optimization, load balancing, fault tolerance, security etc.

4. Conclusion

There are numerous research challenges that still needs to be addressed with respect to cost optimization, load balancing, and fault tolerance. A lot of research is been carried out across the globe. A model or a mechanism which can provision the resources with minimum cost to the consumer will be really beneficial, the model will be more effective if it balances the load among all the instance of the resources and does not deny the service in the event of a failure. Having said that we cannot compromise on other issues such as security, over-provisioning and under-provisioning of resources.

5. References

[1] V.Vinothina, R.Sridaran, and Padmavathi Ganapathi, "A Survey on Resource Allocation Strategies in Cloud Computing" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No.6, pp. 34-56, 2012.

[2] Rajkamal Kaur Grewal & Pushpendra Kumar Pateriya, "A Rule-based Approach for Effective Resource Provisioning in Hybrid Cloud Environment", International Journal of Computer Science and Informatics ISSN (PRINT): 2231 – 5292, Vol-1, Iss-4, 2012.

[3] M.Sasitharagai, A.Padmashree, .Dhanalakshmi and S.Gowri, " Dynamic Resource Management In Large Cloud Environments Using Distributed Gossip Protocol", International Journal of Internet Computing ISSN No: 2231– 6965, VOL- 1, ISSN-4 2012.

[4] 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.

[5] Smita Vijayakumar Qian Zhu and Gagan Agrawal, "Dynamic Resource Provisioning for Data Streaming Applications in a Cloud Environment", 2nd IEEE International Conference on Cloud Computing Technology and Science,2010

[6] Sivadon Chaisiri, Bu-Sung Lee, Member, and Dusit Niyato, "Optimization of Resource Provisioning Cost in Cloud Computing " , IEEE Transactions On Services Computing, VOL. 5, NO. 2, April-June 2012.

[7] Rajkumar Buyya, Saurabh Kumar Garg, and Rodrigo N. Calheiros, "SLA-Oriented Resource Provisioning for Cloud Computing: Challenges, Architecture, and Solutions", 2011 International Conference on Cloud and Service Computing.

[8] Xiaoqiao Meng, Canturk Isci, Jeffrey Kephart, Li Zhang, Eric Bouillet, and Dimitrios Pendarakis, "Efficient Resource Provisioning in Compute Clouds via VM Multiplexing", ICAC'10, June 7–11, 2010, Washington, DC, USA.

[9] Qi Zhang · Lu Cheng · Raouf Boutaba, "Cloud computi ng: state-of-the-art and research challenges", 20 April 2010, The Brazilian Computer Society 2010.

[10] Mohsen Amini Salehi, Bahman Javadi, and Rajkumar Buyya, "Resource Provisioning based on Preempting Virtual Machines in Resource Sharing Environments", Published online in Wiley InterScience (www.interscience.wiley.com).

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

[12] Bahman Javadi , Jemal Abawajyb, and Rajkumar Buyya, "Failure-aware resource provisioning for hybrid Cloud infrastructure", Journal of parallel and distributed computing", Elsevier 72 (2012) 1318–1331.

IJERT

[13] Guofu Feng, Saurabh Garg, Rajkumar Buyya, and Wenzhong Li, "Revenue Maximization Using Adaptive Resource Provisioning in Cloud Computing Environments", 2012 ACM/IEEE 13th International Conference on Grid Computing.

[14] Shivam Nagpal, and Parveen Kumar, " A Study on Adaptive Fault Tolerance in Real Time Cloud Computing", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013, ISSN: 2277 128X.

[15] M Dakshayini, P Jayarekha and H S Guruprasad, "Adaptive Fault Tolerant Service Provisioning [AFTSP] Strategy For Cloud Environment", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) Vol. 3, Issue 1, Mar 2013, 351-360, ISSN 2249-6831.

[16] M.Karthi, and S.Nachiyappan, " Survey on Resources Provisioning in Cloud Systems for Cost Benefits", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 3 March 2013 Page No. 831-835.

IJERT