Aim: - To study Cloud Capacity planning/management , their various approaches, different trends and summarizing them.

Abstract: -

[1] When it comes to resources, cloud computing gives the impression of a limitless pool that can be accessed on demand and with a great deal of leeway. Hidden beneath this facade, however, are hundreds of servers, petabytes of storage, tens of thousands of applications, and millions of users. Due to factors including variable demand, a lack of standardization in resources, conflicting management goals, and a massive footprint, the administration of such systems is far from simple. [1] There has been a lot of discussion about cloud computing recently, but the way most people understand it has changed throughout the years. The idea dates back to 1961, when John McCarthy proposed making computers a public service like electricity and telephones. Recently, in 2001, IBM began working on a project named Oc'eano  to realize a cloud architecture, which is an e-business computing utility infrastructure that can dynamically assign resources to customers as their needs change.

The present concept of clouds is solidly anchored in the industry when in 2006 Amazon’s chief executive, Jeff Bezos, unveiled the two cloud \scomputing products of the corporation, EC2 and S3. Amazon developed these products because it had become evident that maintaining a dependable, scalable infrastructure in a typical multi-datacenter model for an extended period of time would be exceedingly costly and would demand a considerable investment of intellectual capital. They consequently began to offer infrastructural services to other firms, offering to lower their clients’ capital and operational costs while also employing Amazon’s own huge resources in a more efficient and profitable way.

The foundation for cloud computing was laid by earlier computing paradigms including grid computing and utility computing.

Each of these theoretical frameworks provided a key idea that was crucial in laying the groundwork for this new line of thought. As a solution to the problem of needing extremely large amounts of computing power, grids were initially adopted by the scientific community.

Meanwhile, the advent of utility computing brought with it the idea of selling computer power in the form of a service, with customers paying for it on an as-needed basis.

Both paradigms, however, lacked crucial characteristics that would have increased their usefulness and facilitated their widespread adoption, such as adaptability, on-demand availability, or the capability for the short-term acquisition of seemingly endless computing resources. In this paper we will study cloud capacity planning/management , various approaches used, their niche uses.

Introduction

[2] Due to its low maintenance, widespread availability, and great efficiency, cloud services have rapidly become an indispensable IT backbone for numerous modern business initiatives. It has the versatility to offer a wide variety of services over the web, including software as a service, platform as a service, and infrastructure as a service. One type of cloud service, infrastructure as a service (IaaS), attempts to deliver an elastic and unified platform for running virtual machines (VMs) and supporting other IT solutions for enterprise operations. When it comes to properly arranging and assigning the cloud resource in paradigms like IaaS, cloud capacity planning and the instant provisioning of VMs are the key tactics for meeting the diverse demands of customers.

IaaS resource requirements are becoming increasingly unpredictable in comparison to traditional IT services because of the pay-as-you-go nature of modern cloud services and rising usage volumes. In order to strike a good balance between customer happiness and the costs of delivering cloud resources, this poses significant difficulties for cloud capacity planning and the rapid provisioning of virtual machines (VMs).

Cloud Capacity Planing is a straightforward strategy that optimizes the assignment of cloud resource capacity based on the customer demands. However, the elastic and volatile demands bring more difficulties in Taking the time to calculate out how much storage space you'll need in the cloud. However, there is the risk of losing customers and money if demand is underestimated and the resulting influx of requests cannot be handled immediately due to insufficient cloud resources. However, if the requirements are overstated, then the available cloud resources will be underutilized. There will be significant strains on the management of data centers due to the need to invest in costly infrastructures early and the high expense of ongoing maintenance.

Installing and maintaining a sizable number of cloud infrastructures (say, a thousand server racks) might cost a data center several million dollars each year. The purpose of "instant provisioning" of virtual machines is to swiftly supply requested varieties of VMs. As with cloud capacity planning, if you underestimate or overestimate the demand for individual virtual machines (VMs), you risk losing customers and wasting cloud resources. Customers still have to wait for the initiation of services, even though more and more cutting-edge VMs provisioning methodologies let preparing a specific virtual machine in minutes. For these urgent needs, VM clone techniques become a viable option, since they can quickly and easily construct and supply virtual machines. Nonetheless, only if the calendar of customer requests is provided is it possible to reduce the time consumption of simple procedures like VM verification and automatic patching.

Major Concepts Of Cloud Capacity Management

Resource Allocation [1]

The identification of resources that may fulfill the requirements of each service and the allocation of those resources to the services in a way that is compatible with the cloud provider's aims is a fundamental problem in the management of cloud data center resources. To do this, cloud providers use allocation processes whose quality is typically evaluated in terms of a utility function that stands in for the cloud provider's goals . A scheme that maximizes this utility function is the best one for allocating resources. Profit, resource use, energy efficiency, and distributive justice are all possible objectives that can be used to define the utility function.

It is important that allocation procedures be flexible, as application workloads change over time and the total amount of time an application will run is typically unknown when resources are first allocated. Due to the fluid nature of the allocation process, re-consolidation and over-consolidation of resources may prove useful.

Over-consolidation is undertaken to increase resource utilization in circumstances where the allocated resources are not being used to their maximum capacity, while re-consolidation is performed to retain the optimality of the allocation 9 scheme by re-distributing existing assigned resources.

Information on the available cloud resources in terms of virtual machines (VMs), memory, and CPU capacity; application requirements, typically defined in terms of on-demand capacity (expressed as workloads) or static capacity (predefined VMs that are requested only once); and the provider's requirements, also known as objectives, are all necessary inputs for solving resource allocation problems.

Elasticity[1],[3]

It is the elasticity of cloud computing that sets it apart from other types of distributed computing platforms. It is important for cloud customers to be able to dynamically add and remove resources to better match their needs

Monitoring data is used to estimate demand based on the current load, and there are two types of elasticity algorithms: those that adjust the number of instances of a given type of virtual machine (VM), providing horizontal elasticity, and those that adjust the size of the VMs, providing vertical elasticity.

Elasticity has been the subject of extensive research, and numerous methods have been developed to adapt provisioning to fluctuating application needs. These methods include static threshold setting, control theory, queueing theory, and time series analysis.

Approaches to capacity management

**Method 1 Summary**[1],[40]

In this method, we investigate a re-packing strategy for coordinating the trade-offs between horizontal and vertical elasticity decisions for managing the capacity gained by an elastic application in a cost-effective and on-demand manner.

Since horizontal elasticity options only increase the capacity of the presently deployed VMs, the resulting resource set may not be ideal for the aggregated capacity over time, but it can be implemented quickly and at low cost in terms of reconfiguration. On the other side, vertical elasticity decisions need frequent, costly and time-consuming reconfiguration of the deployed resource set, such as VMs On/Offs; yet they retain the optimality of the resource set. Inevitably, the costs and benefits of the two approaches will have to be balanced against one another.

In this method, we look into how the cost-effectiveness of the resource set can be enhanced by scaling in both the number and the size of virtual machines by combining the advantages of vertical and horizontal elasticity. A cost-benefit analysis is offered that takes the current configuration as well as the cost sand durability of a reconfiguration into account to determine the appropriate trade-off between horizontal and vertical scaling. This analysis is then utilized to inform and make a re-packing decision.

We tested several autoscaling methods in tandem with our repackaging method. The effects of adjusting numerous characteristics on re-packing decisions were explored, including workloads, the cost and durability of sreconfigurations, and the type of application being executed. Using a cost-benefit analysis, we were able to determine when and how to replace the suboptimal set with a new ideal set, resulting in a 60% reduction in overall resource utilization costs throughout the lifetime of the program.

**Method 2 Summary** [1],[41]

Method 2 tackles the problem of holistically managing cloud resources to meet the IP's business level objectives. The study takes the top-down strategy shown in Figure 1 to create a centralized cloud resource management platform. This strategy involves delegating management tasks to a network of low-level controllers. Admission, elasticity control, virtual machine placement, and monitoring and fault management are just some of the administrative duties handled by the various autonomous controllers.

The primary issue to solve is the interdependence among the controllers, which arises from the fact that each controller is complicated and serves a unique purpose. As a result, a single person's choice can have far-reaching consequences for the actions of others, diminishing the solution's fitness for the overall goal.

Graphical user interface, diagram

Description automatically generated with medium confidence

The admission controller may be more lenient in allowing services if, for example, the server consolidation level (number of VMs per physical host) were increased or the elasticity control were made more stringent. Thus, some degree of coordination between these actions is required to optimize the overall behavior of the system to attain a Business Level Objective (BLO).

**Method 3 Summary[1][42]**

Thirdly, a P2P resource management paradigm is proposed for use in cloud computing data centers. The primary goal of the research was to provide a computationally feasible resource management solution that can scale in response to increasing numbers of physical servers and requests for new Virtual Machines (VMs). The framework is made up of a group of agents who share common goals and communicate with one another via a P2P system, all while using a gossip protocol to optimize the sharing of information.

Method 3 takes a bottom-up approach, in contrast to Method 2, by breaking the system down into a large number of interconnected parts. These parts each have both functional and managerial responsibilities. The goal of this decomposition is not to isolate each part's responsibilities, but rather to keep their tasks small enough that their autonomic solutions are sufficient.

This method relies on the cooperation of these autonomous components to fulfill the overarching goal. Figure 2 provides a basic overview of the P2P method.

Graphical user interface, application

Description automatically generated

The strategy used primarily addressed the issue of resource allocation.

In order to increase data center utilization and profit, a gossip-based resource allocation protocol was proposed. This protocol would allocate resources in such a way that all active nodes would be heavily utilized while putting all other nodes into an energy-saving mode.

Data center usage, node utilization, hop count, rejection rate, and net profit were some of the metrics used to assess the new protocol's efficacy. The approach illustrates how request propagation and allocation distribution affect the protocol's efficiency. The method's scalability was investigated by testing its performance on varying numbers of servers. Success was obtained, suggesting the method can be scaled to systems with more than 5,000 nodes and 9,000 placement requests arriving throughout the simulation time, provided that policies are implemented to efficiently propagate these requests.

Finally, a re-consolidation procedure is presented for improving the distribution of already-running virtual machines. To better use active nodes, the re-consolidation process was made to rearrange the allocation of data center resources. Node usage can be improved by as much as 10% with re-consolidation. By decreasing power consumption, this improves the data center's bottom line.

**Future Trends**

1. Trying to find better methods of optimization

[1][42]Recently, heuristics have been applied to the resource allocation problem in the form of an optimization problem.

In a nutshell, heuristics can produce near-optimal solutions in a short amount of time and at a minimal computing cost. A higher quality optimized solution, however, necessitates a more thorough assessment of different optimization methods.

2.The scalability of the P2P method needs more analysis.

In Method 3, [1][42] the suggested P2P method scales in terms of both the number of servers and the volume of incoming VM requests. A minimum of 5000 servers and 9000 VM requests were successfully processed, indicating that the framework is very scalable. However, more research is needed to determine the method's viability at bigger stages. There is a need to investigate the system's high scalability in scenarios where the demand for virtual machines (VMs) grows but the available servers do not.

3. Incorporating new features into the P2P framework

Several key cloud features, such as admission control, failure detection and management, and virtual machine (VM) and application placement, are supported by the P2P architecture. As a result, this has helped alleviate some of the issues with VM-level resource placement and distribution.

4.Incorporating additional BLOs into the P2P framework

At now, we structure the allocation problem so as to maximize both data center usage and financial gain. Methods that can account for more nuanced corporate goals, such equality, energy efficiency, and SLA violation reduction, would be very helpful.

For instance, the framework might be modified to accommodate more nuanced limitations, such as anti-affinity and affinity constraints for VM allocation, or consolidation rate and usage limits for controlling data center overbooking.

Capacity planning in Private Cloud

CLOUD MONITORING CHALLENGES[6]

1. **Monitoring Ownership**[6]- Most of the time, the owner of an application that runs on dedicated servers is also responsible for monitoring those servers. They rely on tracking devices to detect any infrastructural issues and the need to track down use data. Using a private cloud, several programs can share a single set of servers. This means that those responsible for deploying and maintaining an application in the cloud must now share responsibility for its monitoring.

Service provider in the cloud computing industry who offers infrastructure services to the owners of applications.

1. **Visualization Of monitored data**[6]- Information like resource use, application health details, changes in application deployment, etc., may all be gleaned from cloud monitoring. It's possible that daily data production will reach the terabyte range. In order to make sense of the complicated and high-volume data that cloud monitoring systems can give, both application owners and cloud infrastructure providers need access to specialized visualization tools[2].

Some common examples of when depiction is needed are given below.

Providers of infrastructure will benefit from having a visual depiction of an application's and its components' locations on a server layout in the event of an issue. Please keep in mind that this is a dynamic diagram and subject to change.

The owner of the infrastructure can benefit from having access to a heatmap of all virtual machines currently being hosted on real servers.

**IBM Smartcloud Monitoring [6]:-**

With IBM SMARTCLOUD MONITORING, you can keep an eye on your cloud infrastructure 24/7. (SCM)Many of the needs discussed above can be met with the help of IBM SmartCloud Monitoring. It's a collection of tried and true IBM Tivoli infrastructure management technologies designed to aid the cloud administrator in keeping tabs on and preparing for the growth of private clouds.SCM's technical capabilities consist of the following [6].Access to the cloud's underlying infrastructure: Including both physical and virtual servers, storage, and network resources, SCM gives holistic, contextual views of virtual environment health and performance.Information on the state of the cloud infrastructure can be accessed through health dashboards, which also serve to display real-time and historical metrics. Here is a sample of the dashboard:

Chart, bar chart

Description automatically generated

-Supports Contingency plans for capacity

-Virtual machines (VMs) placement optimization based on policy to meet application needs.

--The use of performance analytics to determine the optimal number of virtual machines. What follows is an illustration of the suggested sizes.

Graphical user interface, application

Description automatically generated

References:-

[1] Capacity Management Approaches for Compute Clouds Mina Sedaghat

[2] An Effective Classification-based Framework for Predicting Cloud Capacity Demand in Cloud Services Bin Xia, Tao Li, Qifeng Zhou, Qianmu Li, and Hong Zhang

[40] ] M. Sedaghat, F. Hernandez-Rodriguez, and E. Elmroth, “A virtual machine re-packing approach to the horizontal vs. vertical elasticity trade-off for cloud autoscaling,” in Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference, p. 6, ACM, 2013.

[41] M. Sedaghat, F. Hern´andez, and E. Elmroth, “Unifying cloud management: Towards overall governance of business level objectives,” in 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), pp. 591–597, IEEE, 2011.

[42] M. Sedaghat, F. Hern´andez, and E. Elmroth, “Peer to peer resource management for cloud data centers,” Submitted to 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2014.

[6] Monitoring and Capacity planning of private clouds: The challenges and the solution Geetha Adinarayan, India Software Lab, IBM, Bangalore, India