International Journal of Engineering Research & Technology (IJERT)
NCITSF`14 Conference Proceedings
ISSN: 2278-0181

The Comprehensive Study on Saving Energy in Cloud Computing

Manjula L
II Sem, Mtech(CNE)
Dr. Ambedkar Institute of Technology
Bangalore, Karnataka
manjularaghu10@gmail.com

Shylaja B S
Professor& HOD, Dept of ISE
Dr. Ambedkar Institute of Technology
Bangalore, Karnataka
shyla.au@gmail.com

Anushree R
II Sem, Mtech(CNE)
Dr. Ambedkar Institute of Technology
Bangalore, Karnataka
anushreya.r@gmail.com

Abstract- **Cloud computing is an emerging computing technology uses the internet and center remote servers to maintain the data and applications. Energy accounts for a significant fraction of the operational costs of a data center. Designing energy efficient data centers is a major issue in cloud computing. This paper addresses different saving energy techniques like offloading, scaling and right sizing data centers for power saving in cloud computing.**

Keywords- cloud computing, datacenter, energy efficiency, server.

I. INTRODUCTION

Cloud computing is an emerging style of computing where applications, data and resources are provided to users as a services over the web. The services provided may be available globally, low in cost, massively scalable, pay as you grow. Consumers of a service need to care only about what the service does for them and not on how it is implemented. Cloud computing is a model driven methodology that provides configurable computing resources such as servers, networks, storage and applications as and when required with minimum efforts over the internet services.

Cloud computing needs a datacenter which is a facility used to house computer systems and associated components such as telecommunications and storage systems. Data centers have their roots in the huge computer rooms of the early ages of the computing industry. Data centers are tied to locally with specific components including redundant power supplies, redundant communications, environmental controls, security devices etc. Clouds are location independent providing abstracted versions of data center components that are not tied to a specific datacenter's virtual servers, virtual networking and so on. Clouds almost certainly span one or more data centers but, they are not data centers.

Figure 1 shows the typical architecture of data center servers [1]. It consists of a load balancing switch, a set of application servers, and a set of backend storage servers. The front end switch is typically a Layer 7 switch parses application level

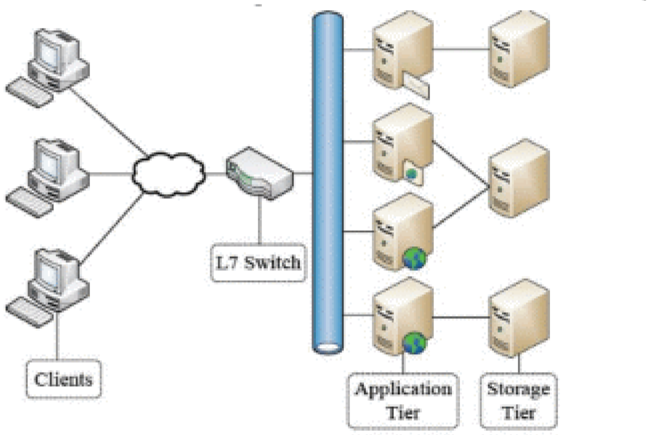

Fig.1 General Architecture of Data centre servers in cloud computing

servers with the corresponding applications running[2]. The switch sometimes runs in a redundant pair for fault tolerance. Each application can run on multiple server machines and the set of their running instances are often managed by some clustering software such as Web Logic [3]. Each server machine can host multiple applications. The applications store their state information in the backend storage servers. It is important that the applications themselves are stateless so that they can be replicated safely. The storage servers may also become overloaded, but the focus of this work is on the application tier [10]. The architecture is a representative of a large set of Internet services hosted in the cloud computing environment [9]

The paper is organized as follows: Section 2 presents the issues in data center. Section 3 includes the different techniques for energy saving and finally Section 4 includes conclusion.

ISSN: 2278-0181

II. ISSUES IN DATA CENTER

Important issues in designing the data centers are redundancy, efficiency, utilization and management. Energy use is a central issue for data centers. Power draw for data centers ranges from a few kW for a rack of servers in a closet to several tens of MW for large facilities. Some facilities have power densities more than 100 times that of a typical office building. For higher power density facilities, electricity costs are a dominant operating expense and account for over 10% of the total cost of ownership (TCO) of a data center. In 2007 the entire information and communication technologies or ICT sector [6] was estimated to be responsible for roughly 2% of global carbon emissions with data centers accounting for 14% of the ICT footprint. The US EPA estimates that servers and data centers are responsible for up to 1.5% of the total US electricity consumption or roughly .5% of US GHG emissions, for 2007[7]. Given a business as usual scenario greenhouse gas emissions from data centers is projected to more than double from 2007 levels by 2020[8].

III. DIFFERENT TECHNIQUES FOR SAVING ENERGY

Energy costs represent a significant fraction of a data center's budget and this fraction is expected to grow as the price of energy increases in coming years. Hence, there is a growing push to improve the energy efficiency of the data centers behind cloud computing Several issues exists for the energy saving in cloud computing namely Scaling, computation offloading, power proportional datacenters and so on. Different examples are illustrated below:

A. Automatic Scaling of Internet Applications for Cloud Computing Services:

Internet applications can benefit from an automatic scaling property where their resource usage can be scaled up and down automatically by the cloud service provider [1]. Each application instance is encapsulated inside a virtual machine (VM) and use virtualization technology to provide fault isolation. It is modeled as a Class Constrained Bin Packing (CCBP) problem where each server is a bin and each class represents an application. The class constraint reflects the practical limit on the number of applications a server can run The architecture of auto scaling system is shown in fig. 2 .Each application instance is encapsulated inside a virtual machine (VM). VMs provide isolation among untrusted users simultaneously.

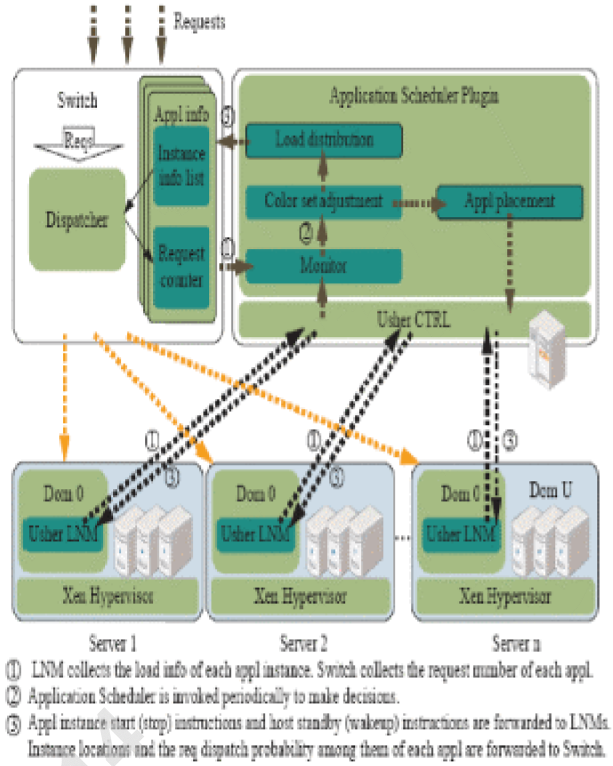

Fig .2 Architecture of Auto scaling System

. Each server in the system runs the Xen hypervisor, which supports a privileged domain 0 and one or more domain U. Each domain U encapsulates an application instance, which is connected to share network storage (i.e., the storage tier). The multiplexing of VMs to PMs (Physical Machines) is managed using the Usher framework.

The main logic of the system is implemented as a set of plug-ins to usher. Each node runs a Usher Local Node Manager (LNM) on domain 0 which keeps track of the set of applications running on that node and the resource usage of each application. A L7 switch is in charge of forwarding requests and responses. The schedule procedure of the system can be described as follows:

• The LNM at each node and the L7 switch collect the application placement, the resource usage of each instance, and the total request number of each application periodically. Then the information is
forwarded to the Usher central controller (Usher CTRL) where the "Application Scheduler" runs.

• The Application Scheduler is invoked periodically to make the load distribution and application instances
The LNM at the node adjusts the local resource allocation of the VMs encapsulating the applications.

Xen can change the CPU allocation among the VMs by adjusting their weights in its CPU scheduler. Memory allocation among the VMs can be adjusted using the ballooning technique. After that the Scheduler notifies the L7 switch of the new configuration including: the list of applications and for each application, the location of its running instances and the probability of request distribution among them

The L7 switch then starts processing Web requests according to the new configuration. The decision interval of the Scheduler depends on how responsive the application demand change. Generally Usher CTRL is a central point of failure. Fortunately, it is not involved in the normal processing paths of user requests. Its failure only disrupts the updates on the load distribution and application placement policy. Incoming user requests can still be processed by the existing application instances. It can be noticed that complicated applications can take a long time (several minutes or much longer) to start and finish all the initializations. Since the application runs inside a VM, its entire running state can be suspended to the disk and then resumed at a later time.

Zhen Et.al [1] results demonstrate that system can improve the throughput by 180% over an open source implementation of Amazon EC2 and restore the normal QoS five times as fast during flash crowds. Large scale simulations demonstrate that it is extremely scalable: the decision time remains under 4 seconds for a system with 10,000 servers and 10,000 applications.

B. When Mobile Terminals Meet the Cloud: Computation Offloading as the Bridge:

The emergence of cloud computing has been dramatically changing the landscape of services for modern computer applications. Offloading computation to the cloud effectively expands the usability of mobile terminals beyond their physical limits, and also greatly extends their battery charging intervals through potential energy savings.

This example[5] provides an overview of computation offloading in mobile cloud computing. The key issues are identified in developing new applications that effectively leverage cloud resources for computation-intensive modules, or migrating such modules in existing applications to the mobile cloud.

Figure 3 shows a conceptual architecture for mobile cloud computing. The mobile terminals access the Internet via WiFi or cellular networks, and coordinate with application servers to locally decide on the offloading strategy. Then mobile terminals will offload the tasks to the cloud accordingly. Upon receiving the requests from mobile terminals or mobile application servers, the cloud Controllers will schedule the tasks on *virtual machines*, which are rented by application service providers, and send back the results. In some occasions, the application servers can also be deployed in the cloud.

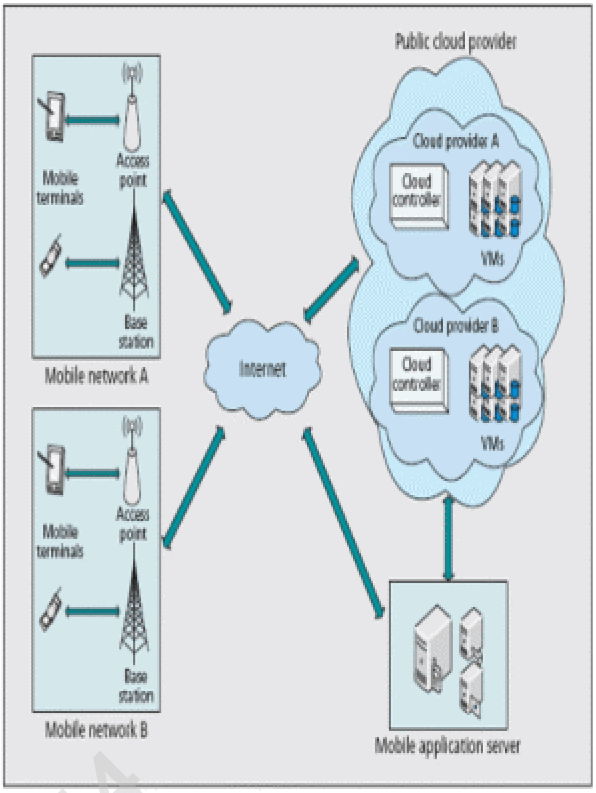

Fig. 3 Architecture for mobile cloud computing

The mobile cloud computing inherits the salient benefits of general cloud computing. In particular, dynamic provisioning allows application providers to handle the variation of service burden in a flexible and fine-grained way rather than over provisioning in advance, which is a waste of money and resources. Application providers can easily expand the system scale to meet increasing user demands. They can also integrate different services through the cloud for efficient management. Virtualization, a key feature in cloud computing, provides isolation and protection for individual virtual machines, which significantly improves system reliability. It also offers higher utilization of infrastructure so that the running costs can be largely reduced. First, the key motivation of offloading for a mobile application must be determined: to save energy, to improve computation performance, or both. The potential offloading gain needs to be well understood. There is no incentive to resort to clouds for a job that can easily and efficiently be executed locally. Even for one that can hardly be executed locally, moving it to the remote cloud may incur a large volume of data transfer, which, although it may not be a severe problem for users with high-speed wired network connections, can largely contradict the benefit for mobile users with their energy-hungry wireless interfaces.

C. Dynamic Right-Sizing for Power-Proportional Data Centers:

Power consumption imposes a significant cost for data centers implementing cloud services[4], yet much of that power is used to maintain excess service capacity during periods of low load. This example investigates how energy can be saved by dynamically "right-sizing" the data center by turning off servers during low load periods and how to achieve that saving via an online algorithm. It proposes a very general model and proves that the optimal offline algorithm for dynamic right-sizing has a simple structure when viewed in reverse to develop a new "lazy" online algorithm, which is proven to be 3-competitive.

Algorithm: Lazy Capacity Provisioning LCP (w)

Let $X^{LCP(w)}=(x_o^{LCP(w)}......x_T^{LCP(w)})$ denotes the vector of active servers under LCP(w). This Vector can be calculated using the following forward recurrence relation:

$$x_T^{LCP(w)}=\begin{cases} 0, & t<0 \\ (x_{T-1}^{LCP(w)}\, xU,w_{xt,w,} & t>1 \end{cases}$$

The algorithm using traces from two real data-center workloads and show that significant cost savings are possible. X^{u}_{t} and X^{l}_{t} can be calculated by solving convex optimization of size ŵ+t. The cost of LCP (w) is optimal and does not require significant computational overhead. Cost and energy savings achievable by dynamic right-sizing are significant. The case studies highlight that if a data center has power mean ratio larger than 3, a cost of toggling a server of less than a few hours of server costs, and less than 40% background load, then the cost savings from dynamic right-sizing can be conservatively estimated at larger than 15%. Thus, even if a data center is currently performing valley filling, it can still achieve significant cost savings by dynamic right-sizing.

IV. CONCLUSION

The issues in designing the data centers are presented and with the increasing tendency of cloud computing the need for energy saving mechanisms also increases. The different aspects of energy saving is presented namely scaling, computational offloading and right sizing the data centers. The Scaling approach provides the scaling resources up and down according to the load. Dynamic offloading provides the energy saving in mobile and Right-sizing data centers provide the off and on of servers according to the load. The energy saving techniques provides the efficient design of data centers.

REFERENCES

[1] Zhen Xiao, *Senior Member, IEEE*, Qi Chen, and Haipeng Luo -Automatic Scaling of Internet Applications for Cloud Computing Services. 0018-9340/12/$31.00 © 2012 IEEE.

[2] "Amazon elastic compute cloud (Amazon EC2),http://aws.amazon.com/ec2/."

[3] A. Cohen, S. Rangarajan, and H. Slye, "On the performance of tcp splicing for url-aware redirection," in *Proc. of the 2nd conference on USENIX Symposium on Internet Technologies and Systems*, 1999.

[4] Minghong Lin, Adam Wierman, Lachlan L. H. Andrew, *Senior Member, IEEE, Member, ACM*, and Eno Thereska -Dynamic Right-Sizing for Power-ProportionalData Centers- IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 21, NO. 5, OCTOBER 2013.

[5] Xiaoqiang Ma, Yuan Zhao, Lei Zhang, and Haiyang Wang, Simon Fraser University Limei Peng, SooChow University -When Mobile Terminals Meet the Cloud:Computation Offloading as the Bridge-0890-8044/13/$25.00 © 2013 IEEE.

[6] "Data Center Energy Consumption Trends". U.S. Department of Energy. Retrieved 2010-06-10.

[7] J Koomey, C. Belady, M. Patterson, A. Santos, K.D. Lange. Assessing Trends Over Time in Performance, Costs, and Energy Use for Servers Released on the web August 17th, 2009.

[8] "Quick Start Guide to Increase Data Center Energy Efficiency". U.S. Department of Energy. Retrieved 2010-06-10.

[9] "Google App Engine, http://code.google.com/appengine/."

[10] "WebLogic,http://www.oracle.com/appserver/weblogic/weblogicsuite.html."