ISSN 2277-3061

A SURVEY ON CLOUD COMPUTING AND ITS BENEFITS

Prabhpreet Kaur [(1)], Monika Sachdeva [(2)]
[(1)] Research Scholar, Department of Computer Science Engineering, SBSSTC, Ferozepur
sandhuprabh36@yahoo.com
[(2)] Associate Professor, Department of Computer Science Engineering, SBSSTC, Ferozepur
monika.sal@rediffmail.com

ABSTRACT

Cloud computing is an increasingly popular paradigm for accessing computing resources. In practice, cloud service providers tend to offer services that can be grouped into three categories: software as a service, platform as a service, and infrastructure as a service. Cloud computing represents a shift away from computing as a product that is purchased, to computing as a service that is delivered to consumers over the internet from large-scale data centers – or 'clouds'. This paper discusses some of the research challenges for cloud computing from an enterprise or organizational perspective, and puts them in context by reviewing the existing body of literature in cloud computing. Various research challenges relating to the following topics are discussed: the organizational changes brought about by cloud computing; the economic and organizational implications of its utility billing model; the security, legal and privacy issues that cloud computing raises. It is important to highlight these research challenges because cloud computing is not simply about a technological improvement of data centers but a fundamental change in how IT is provisioned and used. This type of research has the potential to influence wider adoption of cloud computing in enterprise, and in the consumer market too.

Keywords

Cloud Computing, IAAS, SAAS, PAAS, Load Balancing, Virtual Machine.

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY

Vol. 15, No. 4

www.ijctonline.com, editorijctonline@gmail.com

ISSN 2277-3061

INTRODUCTION TO CLOUD COMPUTING

Cloud computing has become an established paradigm for running services on external infrastructure, where virtually unlimited capacity can be dynamically allocated to suit the current needs of customers and where new instances of a service can be deployed within a short time frame. Although the term Cloud computing has come to include several kinds of technologies offering remote execution and service management, it is used in this paper to denote scalable elastic data center infrastructures offering dynamic and cost-efficient service provisioning. There are many different Cloud computing solutions available, such as Amazon Elastic Compute Cloud [1]. However, different Cloud computing solutions are rarely compatible with each other and this creates a kind of vendor lock-in which is not only limiting to the customer, but also limits the potential of Cloud computing as a whole since separate Cloud computing solutions are unable to interoperate. Grid computing can be seen as one of several predecessors to Cloud computing. Grid computing is often about making large computations using large amounts of resources, whileas Cloud computing is more about making large amounts of resources available to many different applications over a longer period of time. Clouds leverage modern technologies such as virtualization to provide the infrastructure needed to deploy services as utilities. Still, Cloud computing and Grid computing share a lot of the underlying technology and many concepts from Grid computing can be modified and made suitable for Cloud computing as well. Cloud computing is Internet("CLOUD") based development and use of computer technology ("COMPUTING").Cloud computing is a general term for anything that involves delivering hosted services over the Internet. It is used to describe both a platform and type of application. These cloud applications use large data centers and powerful servers that host Web applications and Web services. Anyone with a suitable Internet connection and a standard browser can access a cloud application. [1]

Computing started off with the mainframe era. There were big mainframes and everyone connected to them via “dumb” terminals. This old model of business computing was frustrating for the people sitting at the dumb terminals because they could do only what they were “authorized” to do. They were dependent on the computer administrators to give them permission or to fix their problems. They had no way of staying up to the latest innovations. The personal computer was a rebellion against the tyranny of centralized computing [4] operations. There was a kind of freedom in the use of personal computers. But this was later replaced by server architectures with enterprise servers and others showing up in the industry. This made sure that the computing was done and it did not eat up any of the resources that one had with him. All the computing was performed at servers. Internet grew in the lap of these servers. With cloud computing we have come a full circle. We come back to the centralized computing infrastructure. But this time it is something which can easily be accessed via the internet and something over which we have all the control.

BENEFITS OF CLOUD COMPUTING

- Cloud technology is paid incrementally, saving organizations money.
- Organizations can store more data than on private computer systems.
- No longer do IT personnel need to worry about keeping software up to date.
- Cloud computing offers much more flexibility than past computing methods.
- Employees can access information wherever they are, rather than having to remain at their desks.
- No longer having to worry about constant server updates and other computing issues, government organizations will be free to concentrate on innovation.
- Decoupling and separation of the business service from the infrastructure needed to run it.
- Flexibility to choose multiple vendors that provide reliable and scalable business services, development environments, and infrastructure that can be leveraged out of the box and billed on a metered basis—with no long term contracts.

SERVICE MODELS

There are different types of services are provides by cloud models like: Software as a Service(SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [6] which are deployed as public cloud, private cloud, community cloud and hybrid clouds.

1) Software as a Service (SaaS):- The capability provided to the consumer is to use the some applications which is running on a cloud infrastructure. The applications are accessible from many devices through an interface such as a web browser (e.g., web-based email). The consumer does not control the cloud infrastructure which includes network, and servers, all operating systems, and provides storages. Software-as-a-Service (SaaS) is the broadest market. In this case the provider allows the customer only to use its applications. The software interacts with the user through a user interface. These applications can be anything from web based email, to applications like Twitter or Last.fm. This is the idea that someone can offer you a hosted set of software (running on a platform and infrastructure) that you don't own but pay for some element of utilization - by the user, or some other kind of consumption basis. You don't have to do any development or programming, but you may need to come in and configure the (very flexible, configurable and sometimes customizable) software. You don't have to purchase anything. You just pay for what you use. A SaaS provider typically hosts and manages a given application in their own data center and makes it available to multiple tenants and users over the Web.

Some SaaS providers run on another cloud provider's PaaS or IaaS service offerings. Oracle CRM on Demand, Salesforce.com, and Netsuite are some of the well known SaaS examples.

2) Platform as a Service (PaaS):- Platform-as-a-Service (PaaS) is a set of software and development tools hosted on the provider's servers. Google Apps is one of the most famous Platform-as-a-Service providers. This is the idea that someone can provide the hardware (as in IaaS) plus a certain amount of application software - such as integration into a common set of programming functions or databases as a foundation upon which you can build your application. Platform as a Service (PaaS) is an application development and deployment platform delivered as a service to developers over the Web. It facilitates development and deployment of applications without the cost and complexity of buying and managing the underlying infrastructure, providing all of the facilities required to support the complete life cycle of building and delivering web applications and services entirely. PaaS [5] provides all the resources that are required for implementation of applications and all services completely from the Internet. In this no downloading or installing is required of any software. The capability provided to the consumer is to deploy onto the cloud infrastructure .Consumer uses all the applications by using different programming languages and tools which are provide by the provider. Any consumer has not any control on cloud infrastructure including all networks, servers and operating systems, but has control over the applications which they deployed.

3) Infrastructure as a Service (IaaS):- Infrastructure-as-a-Service (IaaS) provides virtual servers with unique IP addresses and blocks of storage on demand. Customers can pay for exactly the amount of service they use, like for electricity or water, this service is also called utility computing. The capability provided to the consumer is to access all the processing, storage , networks and other many fundamental computing resources . Consumer [5] [6] is able to deploy arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage ,deployed application ,and possibly limited control of select networking components.

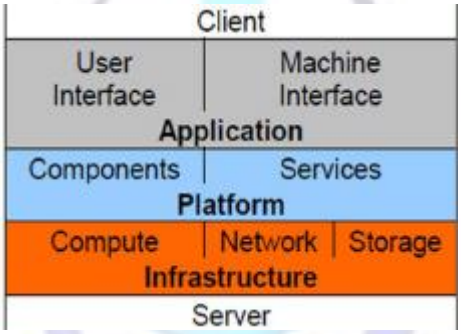

Figure 1. Cloud Computing Stack

DEPLOYMENT MODELS

Depending on infrastructure ownership, there are four deployment models of cloud computing [6].

1) Public Cloud: - Public cloud [9] allows users to access the cloud publicly. It is access by interfaces using internet browsers. Users pay only for that time duration in which they use the service, i.e., pay-per-use. In Public cloud or external cloud resources are dynamically provisioned on a fine-grained, selfservice basis over the Internet, via web applications/web services, from an off-site thirdparty provider who shares resources and bills on a fine-grained utility computing basis.

2) Private Cloud:- A private clouds [10] operation is with in an organization's internal enterprise data center. The main advantage here is that it is very easier to manage security in public cloud. Example of private cloud in our daily life is intranet. Private cloud and internal cloud products claim to "deliver some benefits of cloud computing without the pitfalls", capitalizing on data security, corporate governance, and reliability concerns. They have been criticized on the basis that users "still have to buy, build, and manage them and as such do not benefit from lower upfront capital costs and less hands-on management, essentially.

3) Hybrid Cloud: - It is a combination of public cloud [11] and private cloud. .It provide more secure way to control all data and applications .It allows the party to access information over the internet. It allows the organization to serve its needs in the private cloud and if some occasional need occurs it asks the public cloud for some computing resources. A hybrid cloud environment consisting of multiple internal and/or external providers. It can also describe configurations combining virtual and physical, collocated assets

4) Community Cloud:- When cloud infrastructure construct by many organizations jointly, such cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-party provider or within one of the organizations in the community.

LOAD BALANCING

It is the mechanism of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. It also ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time. Load Balancing [5] is done with the help of load balancers where each incoming request is redirected and is transparent to client who makes the request. Based on predetermined parameters, such as availability or current load, the load balancer uses various scheduling algorithm to determine which server should handle and forwards the request on to the selected server. To make the final determination, the load balancer retrieves information about the candidate server's health and current workload in order to verify its ability to respond to that request. Load balancing solutions can be divided into software-based load balancers and hardware-based load balancers. Hardware-based load balancers are specialized boxes that include Application Specific Integrated Circuits (ASICs) [32] customized for a specific use. They have the ability to handle the high speed network traffic whereas Software-based load balancers run on standard operating systems and standard hardware components.

LOAD BALANCING ALGORITHMS

In order to balance the requests of the resources it is important to recognize a few major goals of load balancing algorithms:

a) **Cost effectiveness**: primary aim is to achieve an overall improvement in system performance at a reasonable cost.

b) **Scalability and flexibility**: the distributed system in which the algorithm is implemented may change in size or topology. So the algorithm must be scalable and flexible enough to allow such changes to be handled easily.

c) **Priority**: prioritization of the resources or jobs need to be done on before hand through the algorithm itself for better service to the important or high prioritized jobs in spite of equal service provision for all the jobs regardless of their origin.

Following load balancing algorithms are currently prevalent in clouds:-

OLB : Opportunistic Load Balancing (OLB) assigns each task, in arbitrary order, to the next machine that is expected to be available, regardless of the task's expected execution time on that machine. The intuition behind OLB is to keep all machines as busy as possible [8, 9].

MET : In contrast to OLB, Minimum Execution Time (MET) assigns each task, in arbitrary order, to the machine with the best expected execution time for that task, regardless of that machine's availability. The motivation behind MET is to give each task to its best machine. This can cause a severe load imbalance across machines.

MCT : Minimum Completion Time (MCT) assigns each task, in arbitrary order, to the machine with the minimum expected completion time for that task. This causes some tasks to be assigned to machines that do not have the minimum execution time for them [1].

Min-min : Min-min heuristic uses minimum completion time (MCT) as a metric, meaning that the task which can be completed the earliest is given priority. This heuristic begins with the set U of all unmapped tasks. Then the set of minimum completion times (M), is found.

Max-Min : The Max-min heuristic is very similar to Min-min and its metric is MCT too. It begins with the set U of all unmapped tasks. Then, the set of minimum completion times (M) is found as mentioned in previous section. Next, the task with the overall maximum completion time from M is selected and assigned to the corresponding machine and the workload of the selected machine will be updated. And finally the newly mapped task is removed from U and the process repeats until all tasks are mapped [1, 9].

LJFR-SJFR : LJFR-SJFR heuristic begins with the set U of all unmapped tasks. Then the set of minimum completion times is found the same as Min-min. Next, the task with the overall minimum completion time from M is considered as the shortest job in the fastest resource (SJFR). Also the task with the overall maximum completion time from M is considered as the longest job in the fastest resource (LJFR). At the beginning, this method assigns the m longest tasks to the m available fastest resources (LJFR). Then this method assigns the shortest task to the fastest resource, and the longest task to the fastest resource alternatively [4, 11].

Sufferage : In this heuristic for each task, the minimum and second minimum completion time are found in the first step. The difference between these two values is defined as the sufferage value. In the second step, the task with the maximum sufferage value is assigned to the corresponding machine with minimum completion time [4, 12].

GREEN COMPUTING

Green computing, also called green technology, is the environmentally sustainable to use of computers and related resources like - monitors, printer, storage devices, networking and communication systems - efficiently and effectively with minimal or no impact on the environment. Green computing whose goals are to reduce the use of hazardous materials, maximize energy efficiency during the product's lifetime, and promote the recyclability or biodegradability of defunct products and factory waste. Conserving resources means less energy is required to produce, use, and dispose of products ,Saving energy and resources saves money .Green computing even includes changing government policy to encourage recycling and lowering energy use by individuals and businesses. Green computing is commonly referred to as

Green IT. The idea is to ensure the least human impact on the environment. Apart from this, it aims to achieve environmental sustainability.

In simple language, green computing is the scientific study of efficient and effective designing, manufacturing, using, disposing, and recycling of computers and computer related products like servers, network systems, communication systems, monitors, USBs, printers, etc. The study uses science to create technologies that help to preserve natural resources and reduce the harmful impact on the environment.

CLOUD SIM

The CloudSim simulation layer provides support for modeling and simulation of virtualized Cloud-based data center environments including dedicated management interfaces for VMs, memory, storage, and bandwidth. The fundamental issues, such as provisioning of hosts to VMs, managing application execution, and monitoring dynamic system state, are handled by this layer. A Cloud provider, who wants to study the efficiency of different policies in allocating its hosts to VMs (VM provisioning), would need to implement his strategies at this layer. Such implementation can be done by programmatically extending the core VM provisioning functionality. There is a clear distinction at this layer related to provisioning of hosts to VMs. A Cloud host can be concurrently allocated to a set of VMs that execute applications based on SaaS provider's defined QoS levels. This layer also exposes the functionalities that a Cloud application developer can extend to perform complex workload profiling and application performance study. The top-most layer in the CloudSim stack is the User Code that exposes basic entities for hosts (number of machines, their specification, and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies. By extending the basic entities given at this layer, a Cloud application developer can perform the following activities: (i) generate a mix of workload request distributions, application configurations; (ii) model Cloud availability scenarios and perform robust tests based on the custom configurations; and (iii) implement custom application provisioning techniques for clouds and their federation.

CONCLUSION

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing provides everything to the user as a service which includes platform as a service, application as a service, infrastructure as a service.

REFERENCES

[1] O.M. Elzeki . "Improved Max-Min Algorithm in Cloud Computing". International Journal of Computer Applications(0975-8887) Volume 50-No.12,july 2012.

[2] "A technical support seminar on cloud computing technology" by Prashant Gupta.

[3] Amandeep Kaur Sidhu. "Analysis of load balancing techniques in cloud computing". International Journal of computers & technology volume 4 No. 2, March-April, 2013, ISSN 2277-3061.

[4] Ektemal Al-Rayis. "Performance Analysis of load balancing Architectures in Cloud computing" 2013 European Modeling Symposium. 978-1-4799-2578-0/13$31.00@2013 IEEE.

[5] Haozheng Ren. "The load balancing Algorithm in cloud computing Environment" 2nd International Conference on computer science and network technology 2012.

[6] Tushar Desai. "A survey of various load balancing techniques and challenges in cloud computing" International Journals of scientific and technology research volume 2. Issue11,Nov2013.

[7] Upendra Bhoi. "Enhanced max-min Task scheduling Algorithm in cloud computing". International Journal of Application or Innovation in Engineering & management(IJAIEM), April 2013.

[8] Klaithem Al Nuaimi, "A survey of load balancing in cloud computing challenges and algorithm". 2012 IEEE second symposium on network cloud computing and applications.

[9] Gytis Vilutis, "Model of load balancing and scheduling in cloud computing". Proceedings of the ITI 2012 34th Int.Conf. on Information Technology Interfaces, June 25-28,Cavat,Croatia.

[10] S. Banerjee, I. Mukherjee and P.K. Mahanti, Cloud Computing Initiative using Modified Ant Colony Framework, World Academy of Science and Technology, 56, pp. 221-224, 2009.

[11] Y. Li, A Bio-inspired Adaptive Job Scheduling Mechanism on a Computational Grid, International Journal of Computer Science and Network Security, 6(3B), pp. 1-7, 2006.

[12] S.C. Wang, K.Q. Yan, W.P. Liao and S.S. Wang, Towards a Load Balancing in a Three-level Cloud Computing Network, Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology, pp. 108-113, 2010.

[13] M. Salehi and H. Deldari, Grid Load Balancing using an Echo System of Intelligent Ants, Proceedings of the 24th IASTED International Conference on Parallel and Distributed Computing and Networks, pp. 47-52, 2006.

[14] M. Dorigo, V. Maniezzo and A. Colorni, Ant System: Optimization by a Colony of Cooperating Agents, IEEE Transactions on Systems, Man, and Cybernetics, PP. 29-41, 1996.

[15] C.W. Chiang, Y.C. Lee, C.N. Lee and T.Y. Chou, Ant Colony Optimization for Task Matching and Scheduling, IEE Proceedings on Computers and Digital Techniques, 153 (6), pp. 373- 380, 2006.

[16] M. Dorigo, M. Birattari and T. Stutzle, Ant Colony Optimization-Artificial Ants as a Computational Intelligence Technique, IEEE Computational Intelligence Magazine, pp. 1- 12. 2006.