Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

The Complete Reference and an Overview of Cloud Computing

Sk. Abdul Kareem* — *IT Department & VLITS, India*
P. Lakshmi Kanth — *IT Department & VLITS, India*
R. Bala Bhargav — *IT Department & VLITS, India*
B.Tarakeswara Rao — *CSE Department & KHIT, India*

Abstract— Cloud computing[1] is a paradigm shift in which computing is moved away from personal computers and even the individual enterprise application server to a cloud of computers. A cloud is a virtualized server pool which can provide the different computing resources of their clients. The data and the services provided reside in massively scalable data centres and can be ubiquitously accessed from any connected device all over the world. Cloud computing is the style of computing where massively scaled IT related capabilities are provided as a service across the internet to multiple external customers and are billed by consumption. Many cloud computing providers have popped up and there is a considerable growth in the usage of this service. Google, Microsoft, Yahoo, IBM and Amazon have started providing cloud computing services. Amazon is the pioneer in this field. Smaller companies like Smug- Mug, which is an online photo hosting site, has used cloud services for the storing all the data and doing some of its services. Cloud Computing is finding use in various areas like web hosting, parallel batch processing, graphics rendering, financial modelling, web crawling, genomics analysis, etc.

Keywords— Big data, Cloud Computing, Hadoop, Map-Reduce, NASDAQ, SmugMug, TimeMachine and Virtualize.

I. INTRODUCTION

The Greek myths give the lowdown of creatures plucked from the surface of the Earth and enshrined as constellations in the night sky. Something similar is happening today in the world of computing. Data and programs are being swept up from desktop PCs and corporate server rooms and installed in “the compute cloud”. In general, there is a shift in the geography of computation.

What is cloud computing exactly? As a beginning here is a definition: “An emerging computer paradigm where data and services reside in massively scalable data centres in the cloud and can be accessed from any connected devices over the internet”.

Like other definitions of topics like these, an understanding of the term cloud computing requires an understanding of various other terms which are closely related to this. While there is a lack of precise scientific definitions for many of these terms, general definitions can be given. Cloud computing [1] is an emerging paradigm in the computer industry where the computing is moved to a cloud of computers. It has become one of the buzz words of the industry. The core concept of cloud computing is, quite simply, that the vast computing resources that we need will reside somewhere out there in the cloud of computers and we’ll connect to them and use them as and when needed. Computing can be described as any activity of using and/or developing computer hardware and software. It includes everything that sits in the bottom layer, i.e. everything from raw compute power to storage capabilities. Cloud computing ties together all these entities and delivers them as a single integrated entity under its own sophisticated management. Cloud is a term used as a metaphor for the wide area networks (like internet) or any such large networked environment. It came partly from the cloud-like symbol used to represent the complexities of the networks in the schematic diagrams. It represents all the complexities of the network which may include everything from cables, routers, servers, data centres and all such other devices.

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 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.

II. CLOUD COMPUTING

Definition for cloud computing[2] can be given as an emerging computer paradigm where data and services reside in massively scalable data centres in the cloud and can be accessed from any connected devices over the internet.

Cloud computing is a way of providing various services on virtual machines allocated on top of a large physical machine pool which resides in the cloud. Cloud computing comes into focus only when we think about what IT has always wanted - a way to increase capacity or add different capabilities to the current setting on the fly without investing in new infrastructure, training new personnel or licensing new software. We have lots of compute power and storage capabilities residing in the distributed environment of the cloud. What cloud computing does is to tie together the capabilities of these resources and make available these resources as a single entity which can be changed to meet the current needs of the user. The basis of cloud computing is to create a set of virtual servers on the available vast resource pool and give it to the clients. Any web enabled device can be used to access the resources through the virtual servers. Based on the computing needs of the client, the infrastructure allotted to the client can be scaled up or down. From a business point of view, cloud computing [2] is a method to address the scalability and availability concerns for large scale applications which involves lesser overhead. Since the resource allocated to the client can be varied based on the needs of the client and can be done without any fuss, the overhead is very low.

One of the key concepts of cloud computing is that processing of 1000 times the data need not be 1000 times harder. As and then the amount of data increases, the cloud computing services can be used to manage the load effectively and make the processing tasks easier. In the era of enterprise servers and personal computers, hardware was the commodity as the main criteria for the processing capabilities depended on the hardware configuration of the server. But with the advent of cloud computing, the commodity has changed to cycles and bytes - i.e. in cloud computing services, the users are charged based on the number of cycles of execution performed or the number of bytes transferred. The hardware or the machines on which the applications run are hidden from the user. The amount of hardware needed for computing is taken care of by the management and the client is charged based on how the application uses these resources.

CHARACTERISTICS OF CLOUD COMPUTING [6]:

1. Self Healing

Any application or any service running in a cloud computing environment has the property of self healing. In case of failure of the application, there is always a hot backup of the application ready to take over without disruption. There are multiple copies of the same application - each copy updating itself regularly so that at times of failure there is at least one copy of the application which can take over without even the slightest change in its running state.

2. Multi-Tenancy

With cloud computing, any application supports multi- tenancy - that is multiple tenants at the same instant of time. The system allows several customers to share the infrastructure allotted to them without any of them being aware of the sharing. This is done by virtualizing the servers on the available machine pool and then allotting the servers to multiple users. This is done in such a way that the privacy of the users or the security of their data is not compromised.

3. Linearly Scalable

Cloud computing services are linearly scalable. The system is able to break down the workloads into pieces and service it across the infrastructure. An exact idea of linear scalability can be obtained from the fact that if one server is able to process say 1000 transactions per second, then two servers can process 2000 transactions per second.

4. Service-Oriented

Cloud computing systems are all service oriented - i.e. the systems are such that they are created out of other discrete services. Many such discrete services which are independent of each other are combined together to form this service. This allows re-use of the different services that are available and that are being created. Using the services that were just created, other such services can be created.

5. SLA Driven

Usually businesses have agreements on the amount of services. Scalability and availability issues cause clients to break these agreements. But cloud computing services are SLA driven such that when the system experiences peaks of load, it will automatically adjust itself so as to comply with the service-level agreements. The services will create additional instances of the applications on more servers so that the load can be easily managed.

6. Virtualized

The applications in cloud computing are fully decoupled from the underlying hardware. The cloud computing environment is a fully virtualized environment.

7. Flexible

Another feature of the cloud computing services is that they are flexible. They can be used to serve a large variety of workload types - varying from small loads of a small consumer application to very heavy loads of a commercial application.

III. NEED FOR CLOUD COMPUTING

What could we do with 1000 times more data and CPU power? One simple question. That's all it took the interviewers to bewilder the confident job applicants at Google. This is a question of relevance because the amount of data that an application handles is increasing day by day and so is the CPU power that one can harness. There are many answers to this question. With this much CPU power, we could scale our businesses to 1000 times more users. Right now we are gathering statistics about every user using an application. With such CPU power at hand, we could monitor every single user click and every user interaction such that we can gather all the statistics about the user. We could improve the recommendation systems of users. We could model better price plan choices. With this CPU power we could simulate the case where we have say 1,00,000 users in the system without any glitches[8].

There are lots of other things we could do with so much CPU power and data capabilities. But what is keeping us back. One of the reasons is the large scale architecture which comes with these are difficult to manage. There may be many different problems with the architecture we have to support. The machines may start failing, the hard drives may crash, the network may go down and many other such hardware problems. The hardware has to be designed such that the architecture is reliable and scalable. This large scale architecture has a very expensive upfront and has high maintenance costs. It requires different resources like machines, power, cooling, etc. The system also cannot scale as and when needed and so is not easily reconfigurable.

The resources are also constrained by the resources. As the applications become large, they become I/O bound. The hard drive access speed becomes a limiting factor. Though the raw CPU power available may not be a factor, the amount of RAM available clearly becomes a factor. This is also limited in this context. If at all the hardware problems are managed very well, there arises the software problems. There may be bugs in the software using this much of data. The workload also demands two important tasks for two completely different people. The software has to be such that it is bug free and has good data processing algorithms to manage all the data.

The cloud computing [3] works on the cloud - so there are large groups of often low-cost servers with specialized connections to spread the data-processing chores among them. Since there are a lot of low-cost servers connected together, there are large pools of resources available. So these offer almost unlimited computing resources. This makes the availability of resources a lesser issue.

The data of the application can also be stored in the cloud. Storage of data in the cloud has many distinct advantages over other storages. One thing is that data is spread evenly through the cloud in such a way that there are multiple copies of the data and there are ways by which failure can be detected and the data can be rebalanced on the fly. The I/O operations become simpler in the cloud such that browsing and searching for something in 25GB or more of data becomes simpler in the cloud, which is nearly impossible to do on a desktop.

The cloud computing [5] applications also provide automatic reconfiguration of the resources based on the service level agreements. When we are using applications out of the cloud, to scale the application with respect to the load is a mundane task because the resources have to be gathered and then provided to the users. If the load on the application is such that it is present only for a small amount of time as compared to the time its working out of the load, but occurs frequently, then scaling of the resources becomes tedious. But when the application is in the cloud, the load can be managed by spreading it to other available nodes by making a copy of the application on to them. This can be reverted once the load goes down. It can be done as and when needed. All these are done automatically such that the resources maintain and manage themselves.

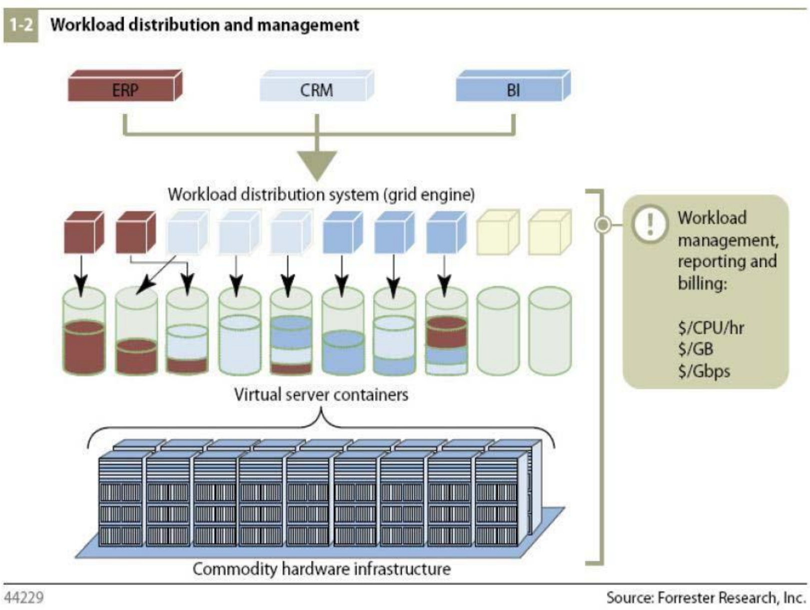

Fig. 1 Work Load Distribution and Management

MAP REDUCE

Map Reduce [7] is a software framework developed at Google in 2003 to support parallel computations over large (multiple petabyte) data sets on clusters of commodity computers. This framework is largely taken from 'map' and 'reduce' functions commonly used in functional programming, although the actual semantics of the framework are not the same. It is a programming model and an associated implementation for processing and generating large data sets. Many of the real world tasks are expressible in this model. MapReduce implementations have been written in C++, Java and other languages. Programs written in this functional style are automatically parallelized and executed on the cloud. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a largely distributed system.

The computation takes a set of input key/value pairs, and produces a set of output key/value pairs. The user of the Map- Reduce library expresses the computation as two functions: Map and Reduce. Map, written by the user, takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key I and passes them to the Reduce function.

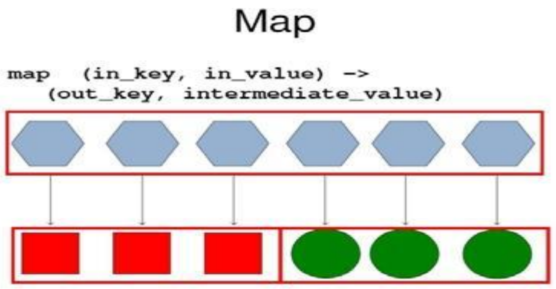

Fig.2 Map Function

The Reduce function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically just zero or one output value is produced per Reduce invocation. The intermediate values are supplied to the user's reduce function via an iterator. This allows us to handle lists of values that are too large to fit in memory.

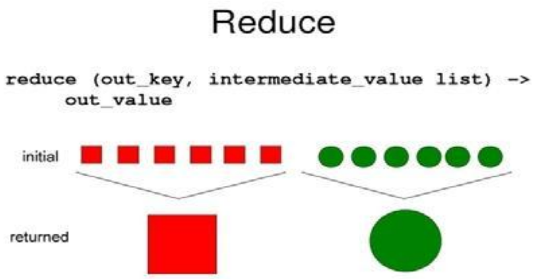

Fig.3 Reduce Function

MapReduce achieves reliability by parcelling out a number of operations on the set of data to each node in the network; each node is expected to report back periodically with completed work and status updates. If a node falls silent for longer than that interval, the master node records the node as dead, and sends out the node's assigned work to other nodes. Individual operations use atomic operations for naming file outputs as a double check to ensure that there are not parallel conflicting threads running; when files are renamed, it is possible to also copy them to another name in addition to the name of the task (allowing for side-effects).

HADOOP

Hadoop [8] is a framework for running applications on large cluster like bigdata built of commodity hardware. The Hadoop framework transparently provides applications both reliability and data motion.

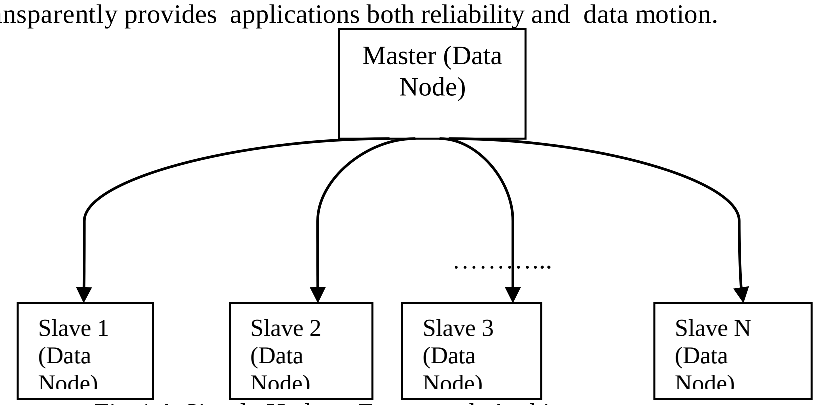

Fig.4 A Simple Hadoop Framework Architecture

The Hadoop MapReduce framework has master/slave architecture. It has a single master server or a jobTracker and several slave servers or taskTrackers, one per node in the cluster. The jobTracker is the point of interaction between the users and the framework. Users submit jobs to the jobTracker, which puts them in a queue of pending jobs and executes them on a first- come first-serve basis. The jobTracker manages the assignment of MapReduce jobs to the taskTrackers.

Hadoop implements the computation paradigm named MapReduce which was explained above. The application is divided into many small fragments of work, each of which may be executed or re-executed on any node in the cluster. In addition, it provides a Hadoop Distributed File System (HDFS) [4] that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. Both MapReduce and the HDFS are designed so that the node failures are automatically handled by the framework. To execute a job, it is submitted to a jobTracker and then executed.

Tasks in each phase are executed in a fault- tolerant manner. If node(s) fail in the middle of a computation the tasks assigned to them are re-distributed among the remaining nodes. Since we are using MapReduce, having many map and reduce tasks enables good load balancing and allows failed tasks to be re-run with smaller runtime overhead.

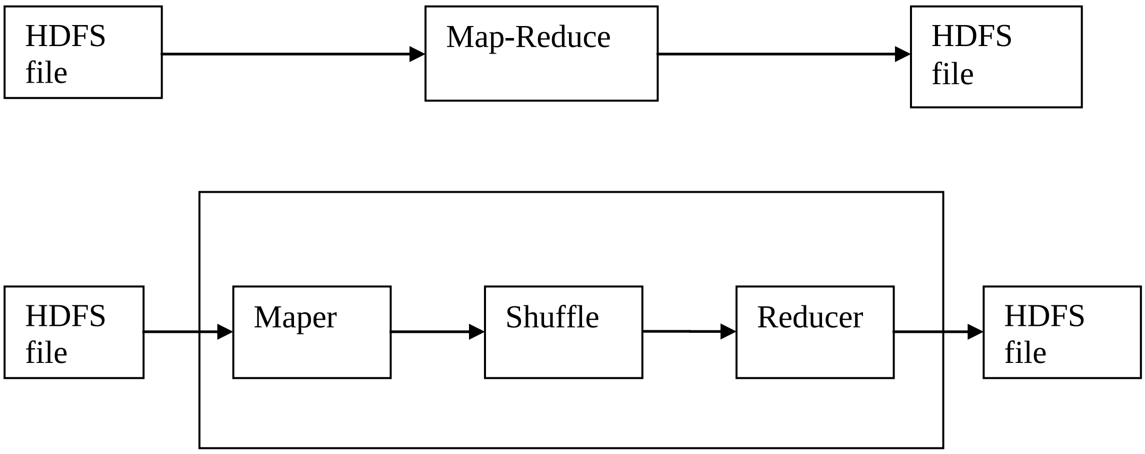

Fig.5 High Level Lexicon for MapReducer

IV. CLOUD COMPUTING IN THE REAL WORLD

Times Machine

Times machine is a New York Times project in which one can read any issue from Volume 1, Number 1 of The New York Daily Times, on September 18, 1851 through to The New York Times of December 30, 1922. They made it such that one can choose a date in history and flip electronically through the pages, displayed with their original look and feel. Here's what they did. They scanned all their public domain articles from 1851 to 1992 into TIFF files. They converted it into PDF files and put them online. Using 100 Linux computers, the job took about 24 hours. Then a coding error was discovered that required the job be rerun. That's when their software team decided that the job of maintaining this much data was too much to do in-house. So they made use of cloud computing services to do the work.

All the content was put in the cloud, in Amazon. They made use of 100 instances of Amazon EC2 and completed the whole work in less than 24 hours. They uploaded all the TIFF files into the cloud and made a program in Hadoop which does the whole job. Using Amazon.com's EC2 computing platform, the Times ran a PDF conversion app that converted that 4TB of TIFF data into 1.5TB of PDF files. The PDF files were such that they were fully searchable. The image manipulation and the search ability of the software were done using cloud computing services.

IBM Google University Academic Initiative

Google and IBM came up with an initiative to advance large-scale distributed computing by providing hardware, software, and services to universities. Their idea was to prepare students "to harness the potential of modern computing systems," the companies will provide universities with hardware, software, and services to advance training in large-scale distributed computing. The two companies aim to reduce the cost of distributed computing research, thereby enabling academic institutions and their students to more easily contribute to this emerging computing paradigm. Eric Schmidt, CEO of Google, said in a statement. "In order to most effectively serve the long-term interests of our users, it is imperative that students are adequately equipped to harness the potential of modern computing systems and for researchers to be able to innovate ways to address emerging problems."

The first university to join the initiative is the University of Washington. Carnegie-Mellon University, MIT, Stanford University, the University of California at Berkeley, and the University of Maryland are also participating in the program. As part of the initiative, Google and IBM are providing a cluster of several hundred computers -- Google's custom servers and IBM Blade Center and System x servers. Over time, the companies expect the cluster to surpass 1,600 processors. The Linux-based servers will run open source software including Xen's virtualization system[5] and Hadoop, an open source implementation of Google's distributed file system that's managed by the Apache Software Foundation. Students working with the cluster will have access to a Creative Commons licensed curriculum for massively parallel computing developed by Google and the University of Washington.

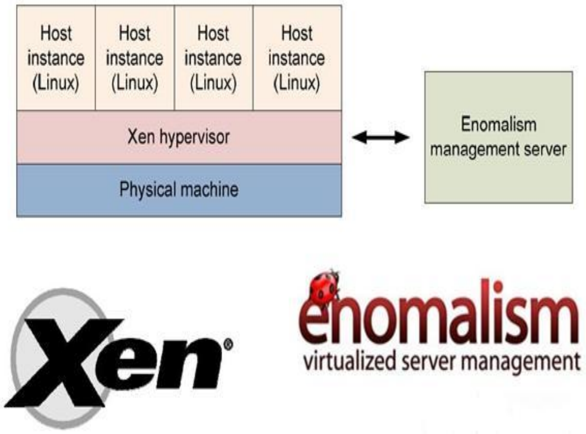

Fig.6 Xen's Virtualized System

SMUGMUG

SmugMug is an online photo hosting application which is fully based on cloud computing services. They don't own any hard drives. All their storage is based in the Amazon S3 instances.

NASDAQ

NASDAQ which had lots of stock and fund data wanted to make extra revenue selling historic data for those stocks and funds. But for this offering, called Market Replay, the company didn't want to worry about optimizing its databases and servers to handle the new load. So it turned to Amazon's S3 service to host the data, and created a lightweight reader app that let users pull in the required data. The traditional approach wouldn't have gotten off the ground economically. NASDAQ took its market data and created flat files for every entity, each holding enough data for a 10-minute replay of the stocks or fund's price changes, on a second-by-second basis. It adds 100,000 files per day to the several million it started with.

V. CONCLUSIONS

Cloud computing [9] is a powerful new abstraction for large scale data processing systems which is scalable, reliable and available. In cloud computing, there are large self-managed server pools available which reduces the overhead and eliminates management headache. Cloud computing services can also grow and shrink according to need. Cloud computing is particularly valuable to small and medium businesses, where effective and affordable IT tools are critical to helping them become more productive without spending lots of money on in- house resources and technical equipment. Also it is a new emerging architecture needed to expand the Internet to become the computing platform of the future.

REFERENCES

[1]. http://www.infoworld.com/article/08/04/07/15FE-cloud- computing-reality_1.html, "*What Cloud Computing Really Means*"
[2]. http://www.spinnakerlabs.com/CloudComputing.pdf *"Welcome to the new era of cloud computing PPT"*
[3]. http://www.johnmwillis.com/ *"Demystifying Clouds" - discusses many players in the cloud space*
[4]. http://www.youtube.com/watch?v=QYzJl0Zrc4M
[5]. http://searchcloudcomputing.techtarget.com/resources
[6]. http://en.wikipedia.org/wiki/Cloud_computing#Research
[7]. http://cdn.ttgtmedia.com/rms/pdf/preparing_for_the_cloud_eguide.pdf
[8]. http://docs.media.bitpipe.com/io_11x/io_110065/item_696401/Migrating%20Applications%20 to%20the%20Cloud_h b_final.pdf
[9]. http://explainingcomputers.com/cloud.html

