IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

An Effective Scheme for Optimal Data Redundancy in Heterogeneous Cloud Infrastructure

Shiny R. M [1] and Dr. Baburaj E [2]

[1]**Computer Science and Engineering, James College of Engineering and Technology, Nagercoil, India**

[2]**Computer Science and Engineering, Sun College of Engineering and Technology, Nagercoil, India**

Abstract

Cloud computing is an emerging technology where user's can gain access to their applications from anywhere through their connected devices. As cloud provides a reliable and transparent storage solution , the data requirements from end users are growing as there is a capability of accessing information from anywhere. Current cloud storage infrastructures are focused on providing users with easy interfaces and high performance services. In this work , a heterogeneous cloud infrastructure is constructed and a precise mechanism is used for measuring the data availability of the nodes. The aim of this work is to measure the data availability of the heterogeneous nodes and to reduce the redundancy and its associated overheads and thus an optimal redundancy can make the system scalable.

Keywords: *Cloud computing, redundancy, data availability, erasure coding*

1. Introduction

Cloud computing [1] [2] is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically-scalable, and managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet. In computational terms, cloud computing is a network of data centers and is described as a powerful, low-cost, and energy efficient approach to future computing. From a sociological standpoint on the other hand, in the cloud, applications are accessible anywhere, anytime, and storage becomes infinite for all intents and purposes and the users can access to powerful applications, platforms, and services delivered over the Internet.
A cloud[3] is a pool of virtualized computer resources. A cloud can host a variety of different workloads, including batch-style back-end jobs and interactive, user-facing applications, allow workloads to be deployed and scaled-out quickly through the rapid provisioning of virtual machines or physical machines, support redundant, self-recovering, highly scalable programming models that allow workloads to recover from many unavoidable hardware/software failures and monitors resource use in real time to enable rebalancing of allocations when needed.

Cloud computing provides computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Virtualization is extremely well suited to a dynamic cloud infrastructure, because it provides important advantages in sharing, manageability, and isolation. Multiple users and applications can share physical resources without affecting one another.

Cloud storage[3] is a model of networked online storage where data is stored on virtualized pools of storage which are generally hosted by third parties. Hosting companies operate large data centers; and people who require their data to be hosted buy or lease storage capacity from them and use it for their storage needs. The data center operators, in the background virtualize the resources according to the requirements of the customer and expose them as storage pools, which the customers can themselves use to store files or data objects. Physically, the resource may span across multiple servers. Cloud storage services may be accessed through a web service Application Programming Interface (API) or through a web based user interface.

Cloud storage[12] is a new distribution model, however, with the potential for economies of scale. Aside from cost, its benefits are outsourced operation, simple, unlimited growth and enterprise features for smaller users - like high availability, security, data protection, etc. At its most basic level, a cloud storage system needs just one data server connected to the Internet. A client (e.g., a computer user subscribing to a cloud storage service) sends copies of files over the Internet to the data server, which then records the information. When the client wishes to retrieve the information, the data can be accessed from the server through a Web-based interface. The server then either sends the files back to the client or allows the client to access and manipulate the files on the server itself.

Cloud storage[13] systems generally rely on hundreds of data servers. Because computers occasionally require maintenance or repair, it's important to store the same information on multiple machines. This is called redundancy. Without redundancy, a cloud storage system couldn't ensure clients that they could access their information at any given time.

Most systems store the same data on servers that use different power supplies. That way, clients can access their data even if one power supply fails. Not all cloud storage clients are worried about running out of storage space. They use cloud storage as a way to create backups of data. If something happens to the client's computer system, the data survives off-site. This expansion ability drives a common expectation that cloud providers have massive storage infrastructures. As a result most cloud storage hardware and software systems developers have also focused on scalability and ease of management. A grid architecture is common, with storage being made up of clusters of individual servers or nodes that are coupled together to present a single storage area or single management point.

The rest of the paper is organized as follows. Section 2 describes the related works. Section 3 presents an overview of the existing system. Section 4 delve into the details of the proposed system. Section 5 presents an experimental evaluation Finally the paper is concluded in Section 6.

2. Related work

Cloud storage services have gained popularity in the last years . These services allow users to store data off-site, masking the complexities of the infrastructure supporting the storage service. To store large amounts of data from thousands of users, cloud storage systems build their services over distributed storage infrastructures, more scalable and more reliable than centralized solutions. Different works have been proposed for distributed data storage infrastructures, some are built on grid infrastructures and some others are built on Peer-To-Peer (P2P) overlays.

Nebulas[8] is an alternative way to construct cloud infrastructures using distributed voluntary resources. Redundancy is the basic technique to provide reliability in storage systems consisting of multiple components. A redundancy scheme defines how the redundant data are produced and maintained. The simplest redundancy scheme is replication, which however suffers from storage inefficiency. Another approach is erasure coding, which provides the same level of reliability as replication using a significantly smaller amount of storage. In the domain of redundant storage[10], the benefit is the reliability of the data storage in spite of failures of the storage components. The ability of a redundancy scheme to be resilient to such failures is usually measured as the probability of a correct reconstruction of a stored object.

Among the existing data redundancy schemes, erasure coding[14] is a widely adopted scheme in existing P2P storage systems. By properly tuning its parameters, erasure codes can minimize the required data redundancy, which reduces both the storage and the network overheads. As a result, the erasure code are optimized ,while deployment reduces its associated overheads. Deng and Wang(2007) have proposed and implemented a heterogeneous storage Grid architecture enabled by Grid service. The storage Grid[15] wraps all distributed and heterogeneous storage resources into grid services to provide transparent, remote, and on demand data access. The architecture provides dynamic and infinite scalability and interoperability of heterogeneous storage resources. Pandey et.al (2010) have presented a Paricle Swarm Optimization (PSO) based heuristic to schedule applications to cloud resources that takes into account both computational cost and data transmission cost. Pamies et.al(2009) have proposed an analytical framework to measure data availability more precisely than existing works. As a result, the erasure code are optimized ,while deployment reduces its associated overheads.

3. Existing System

The existing system are built on distributed infrastructures and the existing redundancy schemes very often assume that the resources are homogenous[10]. Existing storage solutions have considered homogeneous settings, where all nodes are treated equal regarding their on-line/off-line behavior. Although this model is appropriate for commercial clouds, it can be tricky for clouds built of heterogeneous hosts such as nebulas and distributed data centers. Cloud storage services have gained popularity in the last years , these services allow users to store data off-site, masking the complexities of the infrastructure supporting the storage service. To store large amounts of data from thousands of users, cloud storage systems build their services over distributed storage infrastructures, more scalable and more reliable than centralized solutions. Different works have been proposed for distributed data storage infrastructures, some built on grid infrastructures and some others built on P2P overlays. To achieve the desired data availability, all these infrastructures need to store data with redundancy. However, since the use of redundancy increases storage and communication costs.

Existing P2P storage systems[9] do not consider heterogeneity in node availabilities, assuming a unique availability for all nodes in the network. This assumption simplifies the way data availability is measured, but it introduces an error that increases the required redundancy, thereby losing efficiency. Existing P2P storage systems assume homogeneous node availabilities in order to simplify this measurement. This assumption entails efficiency losses when real node availabilities are highly heterogeneous.

4. Proposed System

The proposed system is built on distributed cloud infrastructure with heterogeneous resources. The cloud infrastructure is constructed using volunteer resources. Volunteer resources are attractive several reasons such as scalability ,geographic dispersion, and low cost of deployment[8]. In this system, a precise mechanism is used for measuring the data availability for large set of nodes. This system uses the erasure coding redundancy scheme The main reason is that erasure codes are more flexible than traditional schemes based on replication, a feature that make them very appropriate for heterogeneous platforms. Additionally, they are usually more effective in terms of data overhead than replication. The proposed system reduces the data redundancy which in turn minimizes the communication cost.

A distributed storage system[16] working in its steady state is considered. The number of nodes and the number of stored objects is kept constant. Let N represent the set with all the storage nodes in the system. Then, the storage process stores each data object in a small subset of nodes. Storage process uses an (k, n) erasure code to store objects in the system. Erasure coding is one of the popular redundancy scheme[10]. An erasure code provides redundancy by breaking objects up into smaller fragments and storing the fragments in different places. The key is to recover the data from any combination of a smaller number of those fragments. When recovering data, it is important to check if any fragment is corrupted. It is also important to identify the data to ensure immutability.

Data availability is a probabilistic metric that depends on several parameters. Among all these parameters, the required parameters for measuring d as a function are: the erasure code's parameters k and n, the assignment function g and the set of nodes N. The data availability function $d = D(k; n; g; N)$.This function is used for measuring the data availability of a smaller set of storage nodes around 20.When large set of nodes are used, Monte Carlo approximation is used along with the clustered nodes, A fixed data redundancy is taken and PSO[7] algorithm is used for finding the optimal assignment function .The mean node availability can be increased by storing data only to the high stable nodes in N. The optimization algorithm finds the best assignment.

After finding the data availability and the optimal assignment function , the minimum data availability is found which in turn minimizes the data redundancy. More redundant data is stored in high stable nodes so that it maximizes the data availability which in turn reduces the communication cost. The redundancy ratio of this scheme, r, can be measured . To achieve the desired data availability, all the infrastructures need to store data with redundancy[10]. However, since the use of redundancy increases storage and communication costs, optimizing this redundancy is key to maintain the system scalable. By reducing redundancy, distributed storage systems can reduce storage and communication costs.

5. Experimental Evaluation

This system is implemented in an Open Stack Open Source Cloud Environment. Analysis shows that data redundancy can be reduced upto 80% in heterogeneous cloud environment.

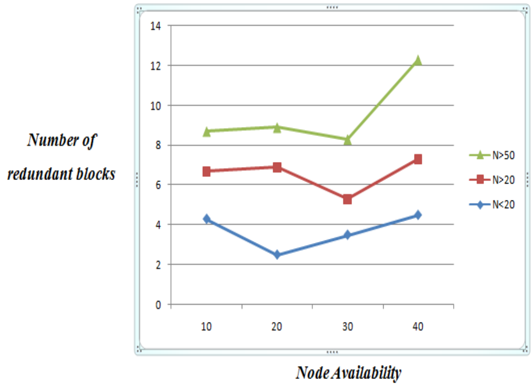

Fig 1:Optimal Redundancy Graph

From the fig, as the number of storage nodes ,N increases the nodes with same data availability can be clustered together and hence using an optimal assignment function and the data availability ,optimal data redundancy can be obtained.

6. Conclusion and Future Work

Cloud computing environments facilitate applications by providing virtualized resources that can be provisioned dynamically. Existing cloud storage services are designed and built on the assumption that all storage back ends constitute a homogenous set of distributed resources. In this work a heterogeneous cloud infrastructure is constructed. This work describes the mechanism to measure the data availability in a precise manner and to optimize the data redundancy so that the system is scalable. This mechanism incurs high computational cost. Future work is to reduce the cost and its associated overheads along with the optimal redundancy.

References

[1]A.Weiss,"Computing in the Clouds,"Networker, ACM, vol. 11(4),Dec 2007,pp 16-25.

[2]I.Foster, Y.Zhao, I.Raicu, S.Y.Lu ,"Cloud Computing and Grid Computing 360-degree compared" Proceedings of Grid Computing Environments Workshop,GCE 2008,IEEE press ,Nov . 2008 .pp .1-10

[3]Jinesh Varia, “Architecting for The Cloud: Best Practices”,Amazon Web Services 2010

[4] Amazon.com, 2009. Amazon S3. <http://aws.amazon.com/s3>.

[5] CleverSafe, 2010, Cleversafe, <http://www.cleversafe.com>.

[6]L. Pamies-Juarez, P. García-López, M. Sánchez-Artigas, "Heterogeneity-aware erasure codes for peer-to-peer storagesystems", in: Proceedings of the 38th IEEE International Conference on Parallel Processing (ICPP), 2009.

[7] Suraj Pandey , Linlin Wu, Mayura Guru, Rajkumar Buyya, “A Particle Swarm Optimization –based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments”, 24th IEEE International Conference on Advanced Information Networking and Applications,2010 ,PP 400-407

[8] Chandra.A., Weissman.J, “Nebulas: Using distributed voluntary resources to build clouds”, Proceedings of the Usenix Workshop on Hot Topics on Cloud Computing (HotCloud'09).,2009

[9]Pamies-Juarez.L, Garca Lopez, Sanchez-Artigas .M,Blas Herrera, “Towards the Design of Optimal Data Redundancy Schemes for Heterogeneous Cloud Storage Infrastructures” Computer Networks, Vol.55,2011,PP-1100-1113

[10] R.Rodrigues ,B.Liskov, “High Availability in DHTS: Erasure Coding vs Replication” , Proceedings of the Fourth Intl. Workshop on Peer to Peer Systems (IPTPS)2005

[11] H.Weatherspoon, J.D.Kubiatowicz,“Erasure coding vs .Replication : A Quantitative Comparison” , Proceedings of the First International Workshop on Peer-To-Peer Systems(IPTPS)2002

[12] Vaquero. L. M, Rodero-Merino. L, Caceres .J, Lindner. M “A break in the clouds: towards a cloud definition”,

SIGCOMM Computer Communication Review, Issue No.1,Volume.39,2009,PP-50-55,

[13] Campbell.R, Gupta. I, Heath.M, Ko.S. Y, Kozuch.M, Kunze. M, Kwan.T, Lai. K, Lee. H. Y, Lyons. M, Milojicic.D, O'Hallaron.D, Soh. Y. C, "Open cirrus cloud computing testbed : Federated data centers for open source systems and services research", Proceedings of the Usenix Workshop on Hot Topics on Cloud Computing (HotCloud'09).,2009

[14] Alessandro Duminuco and Ernst Biersack, "Hierarchical Codes: How to Make Erasure Codes Attractive for Peer-to-Peer Storage Systems", Proceedings of the Eighth International conference on Peer-to-Peer Computing(P2P),2008

[15]Yuhui Deng, Frank Wang , "A Heterogeneous Storage Grid Enabled by Grid Service", ACM SIGOPS Operating Systems Review , 2007 , Issue No .41, Volume 1,PP-7-13

[16] Alexandros G. Dimakis, P. Brighten Godfrey, Martin J. Wainwright and Kannan Ramchandran, "Network Coding for Distributed Storage Systems", 26th IEEE International Conference on Computer Communications (INFOCOM) 2007