IJITKM 0973-4414 Volume-8, Number-2 Jan-Jun 2015 pp. 71-76 available online at www.csjournalss.com

An Overview of Data Replication and Load Balancing Issues in Cloud Computing Environment

Sagar Verma[1], Arun Kumar Yadav[2], Deepak Motwani[3]
[1]M.Tech. Scholar, ITM University, Gwalior, M.P. India
[23]Associate Professor, Department of Comp. Sc. & Engg, ITM University, Gwalior, M.P. India
verma.sagar009@gmail.com, arun26977@gmail.com,deepakmotwani@itmuniversity.ac.in

Abstract: The largest platform for the world to share and store data is based on Cloud computing. The industries are well connected with consumers using IT services. Clients access data and applications through large scale real or virtual networks. The information is accessible between multiple user devices and servicing data centers. The proliferation of clients, buyers and users through multiple devices globally requires a large-scale robust and reliable infrastructure for confirming the confidentiality, integrity and authentication in the global network. The steadfast adherence of Fog computing network provides a medium for authentication, replication and data availability in the middle. The data can reside in cloud servers and provision better proximity through edge networks. The cloud infrastructure demands more resource flexibility on the service side. The survey paper emphasizes on the overloading factor and policies existing for balancing load in cloud servers. The authentication criteria for securing user and critical data in centralized and decentralized hybrid networks are yet to be explored. The issues arising by overloading of Cloud including data explosion and big data challenges are jeopardizing the efficiency of cloud infrastructures. These are inevitably the major issues to resolve in the near future for the virtual computing environments to sustain progressively. The consistency for data and faster service provisioning in terms of response time can make implementation of proximal computing a major asset for maintaining and sharing data for the cloud and the users. The installation of more servers in the Cloud requires hardware resources and finance on a large scale. The ideal system must exhibit minimal delay in service provisioning to the clients and limited transaction failures by having a closer mesh of servers for handling the requests providing data backups, replicas and migrations. That is why, different networking designs and models should be more emphasized on financially and geographically to increase data availability and server deployment for users. The more co-operative Cloud core networks and grid computing can attain a better platform by collectively providing a means to strengthen the virtualized infrastructure for providing versatile services and robustness. The co-operative virtual environments can exhibit few failures in overloading and service provisioning not just for the cloud but the entire mesh of hybrid networks.

Keywords: Cloud, Cloud computing, Response time, Data availability, Data replication, Load balancing, Overloading.

Introduction

Cloud computing is a pool of cohesive platforms facilitating businesses, social networking, marketing and product services .The cloud model houses services for providing information, virtual platforms, infrastructures and softwares frequently on a large scale. The resource investment gradually goes on increasing as the user and service data load on the network increases. All manner of user nodes like wireless sensor nodes, mobile nodes, clients using various devices access the cloud services. The exponentially increasing data related to different users and data centers; the requirement for additional resources is undeniable. The resources in cloud along with the data can be made available in a lesser response time when directed through fog as the middleware. The other issues in cloud services are security issues in providing these services. The fog layer can act as a pre-authentication platform for the user before accessing data directly from edge networks or missing data from cloud. The fog computing supports heterogeneous nodes and more geographical area with less delay in service provisioning .The rate of security threats like phishing and data losses can be reduced by implementing another cloud-like platform providing multi-tenancy and virtualized resources with more proximity towards the client end and a new service infrastructure to implement more security mechanisms in the procedure for improving threat detections and measures.

The Service Level Agreements (SLAs) establishes a relationship between provider and consumer ensuring reliability and consistency of services to the authenticated users. The degree of freedom for clients to use the resources and services requires evaluation and verification of clients to enforce new methods for resource allocation and data migration .The migration of collaborative applications and relevant data are majorly supported among servers. The lack of storage is also forcing migration of data among cloud servers. With rapid advancement in IT industries and technology, sustaining interoperability inside and outside the cloud becomes more challenging. The IT world will require more service vendors, network resources, hardwares, finance and higher level of data optimization and replication to maintain the robustness of the delivery medium.

Load Balancing Metrics

The following characteristics are to be managed efficiently by the load balancing algorithms.

A. Throughput
The system handling large number of tasks in the network should be timely and successfully executed to result in high system throughput.
B. Overhead
The extra cost produced by the load balancing algorithm and approach should be minimal to result in least overhead over the network and individual servers.
C. Fault-Tolerance

The system should keep working even in case of failure. The fault tolerance is the key to achieve system reliability all the time.

D. Migration-Time

The tasks are migrated from one server to another in case the initial server is busy or if the server is overloaded. The time taken by server in data migration should be minimal.

E. Response-Time

The duration of the data delivery and response received should be at it its least to keep the performance ideal like.

F. Resource-Utilization

The cloud infrastructure should be efficient enough to deploy all the resources for handling the processes. Maximum utilization of network resources through a load balancing algorithm is a critical factor in need of optimization.

G. Scalability

The system should be able to sustain the same efficiency in managing network load uniformly with the nodes growing in number.

H. Performance

The performance depends on a collective result of how efficient the cost resource and time management of the algorithm and the network.

An ideal system with high performance at a low cost and resource consumption can be established as a proper standard for developing modern balancing strategies for hybrid networks..The organization’s infrastructural setups handle transactional databases and applications majorly through centralized computing at the core of the network .Moreover, the presentation layer business and behaviors, logics and database functions are established in fog platforms. Implementing such infrastructures increases proximity to the client and faces numerous challenges like application layer challenges, resource provisioning and management, decentralized load balancing mechanisms, deployment and administration of services and availability of data in nearest servers through redundancy, debugging, billing and quantifying the businesses and services. Computing models compatible with J2EE, Microsoft .NET as they can handle specified issues to some extent. The environments for C, PHP and Perl are yet to be presented.

Related Work

Cloud issues presented in [1] “What to migrate” column while discussing cloud issues. The Collaborative applications, IT Management Applications, Business Applications, and Personal Applications collectively each accounted for 25 % of the migrated data in 2008.

The Market Analysis Report presented by IDC for 2012-2015 remarkably highlights the challenges in Big data management with analysis of intelligent economy in the rapidly expanding markets. Big data services and market are stated to increase five times since 2010 demanding large-scale of cloud vendors. Big data houses, R&D, various static and mobile devices, infrastructural data and facilitations provided in large volumes, exceeding 1.8 ZB (1 ZB= 1 billion terabytes) in 2011 estimates. To support cloud load millions of devices in the edge networks will also participate in transfer of data to the client and the cloud.. The demand for administration, management, deployment, and make use of these devices will increase rapidly. The circulation and concentration of data traffic will also be significant rise in edge networks requiring new architectures and computing platforms to support the heterogeneous devices in linked edge networks.[2] The Big Data technology and it's relevant services are estimated by IDC to upsurge by 39.4% compound annual growth rate through 2015.The big data services and infrastructure services accounts for 41% of the total revenue .Replication management scheme for choosing optimum replica number and load balancing is presented in [3] for PC cluster with evaluation of other replication characteristics.

Servers with network and storage sectors accounts for the 30% of worldwide revenue This shows the magnitude of replication resources and strategies required for serving clients and avoid overloading [2,4].Cloud Security Alliance categorized 7 major threats over cloud networks as Shared Technology Issues, Insecure Interfaces and APIs, Data Loss or Leakage, Account or Service Hijacking, Malicious Insiders, Abuse and Nefarious Use of Cloud Computing, Unknown Risk Profiles

The paper aim at performance, availability and administrative attributes of the replication techniques providing scope for academics and research .The multi-tenancy in replication techniques cannot alone solve all the network issues and satisfy customer timely .The stability of system is linked directly with the performance of the load balancing technique. An algorithm is evaluated over certain aspects to determine it's efficiency. No individual node in the network should reach the saturated state in load processing causing system nodes to only circulate the jobs and not processing it [4] [5]. The distribution of load at any time should be uniform over the active nodes, the count of which can frequently vary in the networks .Being a solo technique to provide flexibility in service utilization and sustaining availability to wide network requests through big databases, replication is linked with middleware-based replication for database copies and applications on software level. Master-slave and multi-master replications also described for assisting in increasing database performances [6]. Practical challenges in middleware for RDBMS and SQL are also presented. The details about synchronization, data partitioning for scalability, load balancing and handling conflicts of concurrent transaction failures must be researched upon similarly. [7]

The paper present drawbacks of cluster computing where technology lacks in management and organization of client devices on a large scale .Load balancing is not equally successful in non-parallelized computing than in parallel and distributed systems. The storage is required in abundance, far more than a single server potential. Similarly energy consumption is higher as virtual infrastructure shares distributed resources. Example: Grid computing. Sandboxing technique separates unusual and suspicious sources from the authentic codes, users and programs by inspecting host completely and providing heavy restrictions over security violations. It controls the system architectures due to its restricted operating system .The response time does not get critically influenced by the sudden accumulation of tasks in large numbers [8,12]. Based on distance and processing potential ,the overall response time and throughput holds the maximum priority in selecting a load balancing approach .The overall stability is essential and not just the main data centers state. Even local servers overflow with users in area leading to data loss and shortage of servers. Techniques like round robin are easy to implement due to their static nature but static approach is inadequate for the cloud like environment [9]. B. Radojevic et al. gave algorithm for calculating task execution duration between the two ends of a network on a cloud server along with round robin. Greedy (first-fit) has also been used with round robin to bring better results. The static approaches still were unable to handle all the load of cloud as requests from users can change abruptly with time. In dynamic approach, overload can be distributed to avoid server failure.

In [10], paper presents a min-min load balance algorithm, for controlling resource allocation in dynamic environment. Centralized Load Balancing Approach, E. Hierarchical Load Balancing Approach, Distributed Load Balancing Approach are also discussed to create an efficient hybrid interactive infrastructure .In [11], work is emphasized on distributed load balancing based on similar technique used by honey bees to search for food. And the load balancing technique is so called Honeybee foraging. The algorithms should exhibit higher throughput, minimum response time and maximum resource utilization while some aims at achieving a trade-off amongst all the attributes. [12]

Table 1: Comparing Load Balancing Techniques and Issues

Static	To used fixed values old information is required	Response time Resource utilization Scalability Power consumption and Energy Utilization	Not Scalable User can not alter demands at execution time
Dynamic	Run time statics are required to make the decision	Load estimation. Minimizing the number of migrations. Throughput	Complex ,Time Consuming
Centralized	Load balancing policies are functional in single node.	Threshold strategies Throughput Failure Interaction between central server and processors in network.	No fault tolerance Overloaded central node for decision taking.
Distributed	Load balancing polices are functional in multiple node	Migration, IPC Information exchange criteria, Throughput, Fault tolerance	Algorithm complexity Communication overhead
Hierarchical	Tree data structures are used for decision making for the VM placement, Parent node supervision of other tree nodes.	Threshold strategies Information exchange criteria Selection of nodes at various network layers Failure intensity	Complex, Less fault tolerant

Table 2: Metric Chart Based on Load Balancing Algorithms and Their Overall Efficiency

Metrics	Techniques
Throughput	Round-Robin , Dynamic Round-Robin, PALB , Active Monitoring,FAMLB , Min-Min , Max-Min
Overhead	Round-Robin , Dynamic Round-Robin , PALB , Active Monitoring , FAMLB , Min-Min , Max-Min
Fault tolerance	Dynamic Round-Robin , PALB , FAMLB , Throttled
Migration time	Dynamic Round-Robin , PALB , Active Monitoring , FAMLB , Throttled
Response time	Round-Robin , PALB , Active Monitoring , Min-Min , Max-Min , Throttled
Resource Utilization	Round-Robin , Dynamic Round-Robin , PALB , Active Monitoring , FAMLB , Min-Min , Max-Min , OLB+LB MM , Throttled
Scalability	Round-Robin , PALB , Active Monitoring , FAMLB , Throttled
Performance	Round-Robin , PALB , FAMLB , Min-Min , Max-Min , OLB+LB MM , Throttled

Conclusion

The paper emphasizes on various issues in service provisioning through cloud infrastructures. It highlights the immediate needs for virtualized environment to cope up with the expanding services, applications and platforms, data availability in the networks to maintain robustness and reliability of users in the system. The virtual computing will stay as the basic foundation for network services to end users. The cloud and edge network hierarchies are responsible to keep confidentiality and integrity of the users and make resources and data available through continuous and consistent delivery. With right strategies and administration the tie up should collectively exhibit better response time and reduce costs for IT industries. The edge networks can be significantly supportive in shouldering the big data load and network load on rise, with the cloud. The overhead and overall cost will always be easily handled and controlled if the data and services are facilitated almost though its point of origin rather than a distant centralized node. This also takes less bandwidth and processing to gather, analyze, store and provide data by IoT (Internet Of Things) endpoints and servers. The load of IT sectors and applications can be lowered with using additional distributing devices and processes like routers, OOB management, dynamic resource allocation and strategies for minimal delay in services to end users. With uniform and even distribution of work load between both the networks based on policies for what can be handled through edge networks and what must only be handle through cloud data centers, the hierarchy can provide solutions to many upcoming and existing challenges.

There are many loads balancing algorithm such as Round Robin algorithm, Throttled algorithm, Equally Spread Current Execution Algorithm, Ant Colony algorithm. Randles et.al.[4]giving the comparative analysis by checking cost and performance.

Future Work

The edge networking through its infrastructure can increase scalability in resources and allocate them dynamically to the applicants. There are demands and challenges for developing the distributed cluster of servers in global computing and marketing. It also brings forward the benefits of better end user performance, high service and data availability, improving bandwidth strains, better time-to-market.

The load balancing is evaluated over following metrics: Throughput, Overhead, Fault Tolerance, Response time, Resource Utilization and Scalability .Some techniques like event driven, LBVS, Server-based LB for Internet distributed Services and Vector dot are in use. The ideal load balancing strategy satisfying all the metric values flawlessly is yet to be developed.

The goals are collectively required to be optimized for a load balancing idea to work ideally.

- Uniform load distribution over network servers.
- Provide data redundancy as backup for failures in data access.
- Optimized Routing of load in server infrastructure
- Efficient resource management
- Minimum response time

The strategies of load balancing in the cloud infrastructure are required to reduce the response time and imply a reliable process management system for successfully administering the resources and keeping the cost low. The model should carry good techniques for replicating data through the servers and reduce chances of transaction failure. The management for using replication to kill system failures and switch logic for selection of various load balancing strategies based on the varying characteristics of different cloud networks in the global cloud environment. Techniques like Sandboxing are efficient and necessary to inhibit no interferences between two end users and make it simpler for user to get data immediately while updations and operations in network exhibit no delays. The network system and administration is to be designed such that the infrastructure can be compatible with dynamic arrays of computing platform. The network should eventually become independent with load overloading and failures in resource and data provisioning. The servers in the edge networks ideally can handle requests from all the nodes without failure while cloud handles the only portion of data important to user business and other priorities like integrity and confidentiality. The authentication can take place in the edge networks assisting cloud in sharing data requests and resources. As future work an algorithm for managing load and data replication through edge networks is presented as a solution for cloud overloading and data access failures with better response time.

References

[1]. Kuyoro S.O. Ibikunle F. Awodele O. “Cloud Computing Security Issues and Challenges” International Journal of Computer Networks (IJCN), Volume (3) : Issue (5) : 2011

[2]. Bernd Grobauer, Tobias Walloschek, and Elmar Stöcker,Siemens “Understanding Cloud Computing Vulnerabilities” IEEE Computer Society“ June 2012

[3]. Julia Myint Thinn Thu Naing MANAGEMENT OF DATA REPLICATION FOR PC CLUSTER BASED CLOUD STORAGE SYSTEM International Journal on Cloud Computing: Services and Architecture (IJCCSA), Vol.1, No.3, November 2011 http://airccse.org/journal/ijccsa/papers/1311ccsa03.pdf

[4]. R. M. Bryant and R. A. Finkel, "A Stable Distributed SchedulingAlgorithm," in Proc. 2nd Int. Conf. Dist. Comp., pp. 341-323, April1981.

[5]. D.L. Eager, E.D. Lazowski, and J. Zahorjan, "Adaptive Load Sharing in Homogeneous Distributed Systems," IEEE Trans. Software Eng., vol. SE-12, no. 5, pp. 662-675, May 1986

[6]. Emmanuel Cecchet George Candea Anastasia Ailamaki “Middleware-based Database Replication: The Gaps Between Theory and Practice” *SIGMOD’08*, June 9–12, 2008, Vancouver, BC, Canada

[7]. http://www.ijetae.com/files/Volume4Issue10/IJETAE_1014_64.pdf *WWW*, May, 2004, New York,

[8]. T. L. Casavant, "A Taxonomy of Scheduling in General-Purpose Distributed Computing Systems," IEEE Trans. Software Eng., vol 14, no. 2, pp 141-154, February 1988.

[9]. Z. Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid, “Availability and load balancing in cloud computing,” in International Conference on Computer and Software Modeling, Singapore,

[10]. A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. H. Epema, “Performance analysis of cloud computing services for many-tasks scientific computing,” Parallel and Distributed Systems, IEEE Transactions on, vol. 22, no. 6, pp. 931–945, 2011.

[11].P. Deshmukh and K. Pamu, “Applying load balancing: A dynamic approach,” International Journal, vol. 2, no. 6, 2012.

[12].Comparative study on load balancing technique in cloud computing http://arxiv.org/ftp/arxiv/papers/1403/1403.6918.pdf