International Journal of scientific research and management
(IJSRM) ||Volume||2||Issue||5 ||Pages|| 815-824 ||2014||
Website: www.ijsrm.in ISSN (e): 2321-3418

A Load Balancing Based Cloud Computing Techniques and Challenges

Vikas Kumar[1] and Shiva Prakash[2]

Department of Computer Science & Engineering
Madan Mohan Malaviya University of Technology, Gorakhpur
[1]vikas.vkcool@gmail.com
[2]shiva_pkec@yahoo.com

Abstract **: Cloud Computing refers to the use and access of multiple server based computational resources via a digital network(WAN).Cloud users may access the resources using computer note book, pad computer, smart phone, or other device. In cloud computing applications are provided and managed by the cloud server and data is also stored remotely in cloud configuration. As Cloud Computing is growing rapidly and clients are demanding more services and better results, load balancing for the Cloud has become a very interesting and important research area. Load balancing 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. Here in this paper we have discussed many different load balancing techniques used to solve the issue in cloud computing environment. This paper presents various approaches given by the researchers using the load balancing techniques.**

***Keywords*-Cloud computing, Virtualization, Load Balancing, Fault Tolerance, Load Balancer.**

1. Introduction:

Cloud computing is the use of the pooled computing resources accessible over Internet. Computing resources can be hardware or software. Cloud derives its name from the cloud shaped symbol representing Internet, as it is used as an abstraction for its complex infrastructure. It provides services as per requirement. It allows user to customize, configure, and deploy cloud services. It offers services as per payment. Cloud provides resources over Internet using virtualization technology, multi-tenancy, web services, etc. Virtualization provides abstraction of independent hardware access to each virtual machine. Multi-tenancy allows the same software platform to be shared by multiple applications. Multi-tenancy is important for developing software as a service application. Applications communicate over the Internet using web services [1].

Load balancing is one of the central issues in cloud computing. It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve a high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utility of the system.

There are four deployment models of cloud.

- **Public Cloud**: Public cloud makes services (such as computing, storage, application, etc.) available to general public. These services may be free or offered as payment as per usage. Major public cloud providers are Amazon, Google, Microsoft, etc.
- **Private Cloud:** Private cloud is a cloud infrastructure operated only for a single organization. It is not available to general public.
- **Community Cloud**: Community cloud shared infrastructure between several organization with common concerns such as compliance, jurisdiction, etc.
- **Hybrid Cloud:** Hybrid cloud is a combination of two or more clouds (public, private, or community).

The rest of the paper is organised as follows. In section 1 there is a brief description of cloud computing and cloud infrastructure. Section 2 states the taxonomy in which the service models of the cloud are stated and the technologies related to cloud computing are described, followed by the techniques used in the load balancing. Section 3 describes the various strategies given by the researchers using the load balancing techniques. Section 4 presents a comparative study of the strategies of load balancing given by various researchers. In section 5 the research challenges of cloud computing are stated which describes the scope in which more research has to be made and finally, section 6 presents conclusion and future work.

2. Taxonomy

2.1 Service Models of Cloud

There are four service models of cloud: Infrastructure as a Service, Platform as a Service, Software as a Service, and Network as a Service. Figure 1 demonstrates the abstraction level of services. Software as a service is taken place at the top. From top to bottom services are more fine grained i.e., more access control to the resources [2].

- **Infrastructure as a Service (IaaS):** Cloud provider offer computers as virtual machines and other resources. Virtual machines are run as a guest by a hypervisor such as Xen or KVM. Other resources in IaaS could include images in a virtual machine image library, raw (block storage), file based storage, firewalls, load balances, IP addresses, virtual local area network (VLANs) and

software bundles. IaaS cloud providers supply these resources on demand from their data centers. For wide area connectivity, the Internet can be used. Cloud provides a hosting environment that does not limit an application to a specific set of resources. To deploy their applications, cloud user install operating system image on the machine as well as their application software. In this model, cloud user is responsible for maintaining the operating system and application software.

- **Platform as a Service (PaaS):** Cloud providers provide a computing platform typically including operating system, programming language execution environment (such as Java, Python, Go), database, and web server. Application developers can develop and run their software on a cloud platform. Open source implementation for PaaS are cloud foundry, open shift origin.

- **Software as Service:** In this model, cloud providers install and operate application software in the cloud and cloud users access the software from browser/client interface. Some cloud applications support specific client software dedicated to these applications (e.g., virtual desktop client, email client, etc.).Elasticity makes a cloud applications different from other applications. This can be achieved by cloning tasks onto multiple virtual machines at run time. To accommodate a large number of cloud users, cloud applications can be multi-tenant.

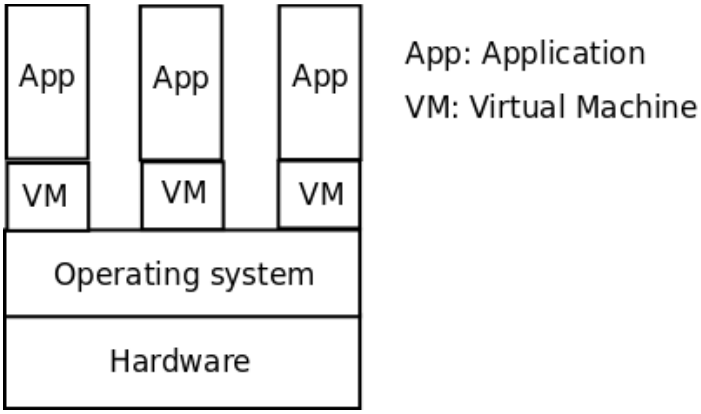
- **Network as a Service:** Cloud service where the capability provided to the cloud service user is to use network connectivity services. NaaS involves the optimization of resource allocations by considering network and computing resources. NaaS services include flexible and extended VPN, and bandwidth on demand.

2.2 Related Technologies

Cloud computing typically has characteristics of all these technologies [3]:
a. Grid computing
b. Virtualization
c. Utility Computing
d. Autonomic Computing
A quick overview of these technologies is given here.

Grid Computing- Grid Computing involves a network of computers that are utilized together to gain large supercomputing type computing resources. Using this network of computerslarge and complex computing operations can be performed. In grid computing these network of computers may be present in different locations. A famous Grid Computing project is Folding@Home. The project involves utilizing unused computing powers of thousands of computers to perform a complex scientificproblem. The goal of the project is "to understand protein folding, mis-folding, and related diseases".

Virtualization- In general sense, a traditional multiprogramming operating system, such as Linux is also a form of virtualization. Linux allows each user process to access system resources without interfering to other processes. The abstraction provided to each process is the set of operating system, system calls and hardware instructions set accessible to user level processes. User mode linux offer a more complete virtual abstraction where each user is not even aware of other user's processes [1]. At a higher level of abstraction are virtual machines based on high level language, such as the Java Virtual Machine (JVM). It runs as an operating system process, but provides a system independent abstraction of the machine to an application written in Java language. Abstraction at the OS system call layer or higher are called process virtual machines as shown in figure:



Utility Computing- Utility Computing defines a "pay-per-use" model for using computing services. In utility computing, billing model of computing resources is similar to how utilities likeelectricity are traditionally billed. When we procure electricity from a vendor, the initial cost required is minimal. Based upon the usage of electricity, electricity companies bills the customer (typically monthly). In utility computing billing is done using a similar protocol. Various billing models are being explored. A few common ones are:
1. Billing per user count. As an example if an organization of 100 people uses Google's gmail or Microsoft Live as their internal email system with email residing on servers in the cloud, Google/Microsoft may bill the organization on per user basis.
2. Billing per Gigabyte. If an organization is using Amazon to host their data on the cloud, Amazon may bill the organization on the disk space usage.
3. Billing per hour/day. As an example a user may pay for usage of virtual servers by time utilized in hours.

2.3 Techniques Used In Cloud Computing

2.3.1 Load balancing

Load balancing is the pre requirements for increasing the cloud performance and for completely utilizing the resources. Load balancing is centralized or decentralized. Load Balancing algorithms are used for implementing. Several load balancing algorithm are introduced like round robin algorithm a mining improvement in the performance. The only differences with this algorithm are in their complicity. The effect of the algorithm depends on the architectural designs of the clouds [4]. Today cloud computing is a set of several data centres which are sliced into virtual servers and located at different geographical location for providing services to clients. The objective of

paper is to suggest load balancing for such virtual servers for higher performance rate.
The paper describes about three load balancing algorithms which are Round robin algorithm, equally spread current execution load and Throttled Load balancing.

- **Round Robin**: Round robin use the time slicing mechanism. The name of the algorithm suggests that it works in the round manner where each node is allotted with a time slice and has to wait for their turn. The time is divided and interval is allotted to each node. Each node is allotted with a time slice in which they have to perform their task. The complicity of this algorithm is less compared to the other two algorithms. An open source simulation performed the algorithm software know as cloud analyst, this algorithm is the default algorithm used in the simulation. This algorithm simply allots the job in round robin fashion which doesn't consider the load on different machines.

- **Equally spread current execution load**: This algorithm requires a load balancer which monitors the jobs which are asked for execution. The task of load balancer is to queue up the jobs and hand over them to different virtual machines. The balancer looks over the queue frequently for new jobs and then allots them to the list of free virtual server. The balance also maintains the list of task allotted to virtual servers, which helps them to identify that which virtual machines are free and need to be allotted with new jobs. The experimental work for this algorithm is performed using the cloud analyst simulation. The name suggests about this algorithm that it work on equally spreading the execution load on different virtual machine.

- **Throttled Load balancing**: The Throttled algorithm work by finding the appropriate virtual machine for assigning a particular job. The job manager is having a list of all virtual machines, using this indexed list, it allot the desire job to the appropriate machine. If the job is well suited for a particular machine than that job is, assign to the appropriate machine. If no virtual machines are available to accept jobs then the job manager waits for the client request and takes the job in queue for fast processing.

2.3.2 Fault Tolerance

- **Reactive Fault Tolerance**

Reactive fault tolerance policies reduce the effect of failures on application execution when the failure effectively occurs. There are various techniques which are based on these policies like Checkpoint/Restart, Replay and Retry and so on.
Check pointing/ Restart - When a task fails, it is allowed to be restarted from the recently checked pointed state rather than from the beginning. It is an efficient task level fault tolerance technique for long running applications [5].
Replication - Various task replicas are run on different resources, for the execution to succeed till the entire replicated task is not crashed. It can be implemented using tools like HAProxy, Hadoop and Amazon Ec2 etc.
Job Migration - During failure of any task, it can be migrated to another machine. This technique can be implemented by using HAProxy.
SGuard- It is less disruptive to normal stream processing and makes more resources available. SGuard is based on rollback recovery [6] and can be implemented in HADOOP, Amazon EC2.
Retry - It is the simplest task level technique that retries the failed task on the same cloud resource.
Task Resubmission - It is the most widely used fault tolerance technique in current scientific workflow systems. Whenever a failed task is detected, it is resubmitted either to the same or to a different resource at runtime. User defined exception handling-In this user specifies the particular treatment of a task failure for workflows.
Rescue workflow - This technique [7] allows the workflow to continue even if the task fails until it becomes impossible to move forward without catering the failed task.

- **Proactive Fault Tolerance**

The principle of proactive fault tolerance policies is to avoid recovery from faults, errors and failures by predicting them and proactively replace the suspected components other working components. Some of the techniques which are based on these policies are Pre-emptive migration, Software Rejuvenation etc.
Software Rejuvenation - It is a technique that designs the system for periodic reboots. It restarts the system with clean state [8].
Proactive Fault Tolerance using Self-Healing - When multiple instances of an application are running on multiple virtual machines, it automatically handles failure of application instances.
Proactive Fault Tolerance using Pre-emptive Migration – Pre-emptive Migration relies on a feedback-loop control mechanism where application is constantly monitored and analyzed.

2.3.3 Task Scheduling

In the cloud computing environment, task scheduling and resource assignment have been unified managed by providers through virtualized technology. They have been used to hide and complete users' tasks transparently. Task scheduling becomes more complex because of the transparent and dynamic flexibility of cloud computing system, and the different needs for recourses of different applications. Task scheduling strategies only focus on equity or efficiency will increase the cost of time, space, throughput and improve the quality of service of the entire cloud computing at the same time. The characteristics of the task scheduling in the cloud computing environment are as follows:
1. Task scheduling caters to a unified resources platform.
As cloud computing using the virtualized technology, we abstracting the underlying physical resources (all types of hosts, workstations or even PC, etc.) as a unified resource pool and shielding heterogeneous, supply the upper use. It mainly distributes in a large number of distributed computers and supply the use of resources in the form of a data center.
2. Task scheduling is global centralized.
As cloud computing is a computing model which supply the centralized resource by the mirror service to multiple distributed applications, and this mirroring deployment can make heterogeneous procedures executing of interoperate become easier, which used to be difficult to deal with. Therefore, virtualized technology and mirroring services make the task scheduling of cloud computing achieve a global centralized scheduling.

3. Each node in the cloud is independent. In cloud computing, the internal scheduling of every cloud node is autonomous, and the schedulers in the cloud will not interfere with the scheduling policy of these nodes.
4. The scalability of task scheduling. The scale of resources supply from cloud provider may be limited in early stages. With the addition of a variety of computing resources, the size of the abstract virtual resources may become large and the application demand continue increasing. In the cloud, task scheduling must meet the scalability features, so that the throughput of the task scheduling in the cloud may not be too low.
5. Task scheduling can be dynamically self-adaptive. Expanding and shrinking applications in the cloud may be necessary depend on the requirement. The virtual computing resources in cloud system may also expand or shrink at the same time. The resources are constantly changing, some resources may fails, new resources may join in the clouds or restart.
6. The set of task scheduling. Task scheduling is divided into two parts: one is used as a unified resource pool scheduling, and primarily responsible for the scheduling of applications and cloud API; the other is for the unified port resource scheduling in the cloud, for example, Map Reduce task scheduling. However, each scheduling consists of two two-way process: scheduler leases resource from cloud, scheduler call backs the requested resources after use. The former process is scheduling strategy and the latter one is callback strategy [9,10]. The combination of the scheduling and callback resource strategy is the set of task scheduling [11].

3.1 Literature Review

1) Jasmin James et al. [23] analysed various VM load balancing algorithms. Then, the author has proposed a new load balancing algorithm and implemented it for an IaaS framework in Simulated cloud computing environment i.e."Weighted Active Monitoring Load Balancing Algorithm".
In the VM load balancing algorithm the author has proposed an approach in which individual application services is assigned varying (different) amount of the available processing power of VMs This is because- in the real world, it's not necessary all the VMs in a DataCenter has fixed amount of processing powers but it can vary with different computing nodes at different ends. And then to these VMs of different processing powers, the tasks/requests (application services) are assigned or allocated to the most powerful VM and then to the lowest and so on. They are given the required priority weights. Hence, the performance parameters such as overall response time and data processing time are optimized.
The author has concluded that the proposed load balancing algorithm has optimized the given performance parameters such as response time and data processing time, giving an efficient VM load balancing algorithm.

2) JaspreetKaur et al. [24] have defined the problem as, the random arrival of load in such a environment can cause some server to be heavily loaded while other server is idle or only lightly loaded .Equally load distributing improves performance by transferring load from heavily loaded server. Thus, the author has provided the solution based on scheduling algorithm ECSE(Equally Spread Current Execution)load. The author has then compared the performance result of ECSE with round robin scheduling to estimate response time, processing time.
In the algorithm proposed by the author the active VM load balancer finds the next available VM. The load balancer checks that for all current allocation count is less than maximum length of VM list in order to allocate the VM. If available VM is not allocated then a new one is created. And at last active load is counted on each VM and id of those VM is returned which is having least load.
The author concludes his paper after simulating the algorithm on CloudSim. The ESCE algorithm dynamically allocates the resources to the job in queue leading reduced cost in data transfer and virtual machine formation. The simulation result shows overall time cost results and comparison of load balancing algorithms. ESCE load balancing provide better results as compared to round robin on closet data centre, optimize response time.

3) NyandeepSran et al. [25] developed a Load Balancer Algorithm that controls the flow of payload based on the safety thresholds, which may be static or dynamic in nature, depending on the available machines and bandwidth as well. The author has analyzed the existing algorithms of Load Balancing such as Round Robin, Throttled, Equally Spread and Biased Random Sampling and has proposed a new algorithm which will meliorate the existing Load Balancing Approach, by decreasing the overall requesting time and processing time as compared to the existing algorithms and hence will decrease the cost which is proved through rigorous simulation study. The Proposed Algorithm will also provide security to the data in cloud during Load Balancing process by using Zero Proof Algorithm.
In the proposed approach the author is working on VM migration policy by giving priority to VM's on the basis of resources available to it. Firstly, the algorithm will check that whether the CPU utilization of VM(Virtual Machine) is equal to, greater than or less than 80%. The value has been chosen in order to prevent the machine from being overloaded.
It is clear from the results that the proposed method in the paper is able to balance the load to a greater extent than the analyzed algorithms. A comparative study is done in this thesis with the Proposed Algorithm. This Load Balancing algorithm aims at providing dynamic, on-demand, balance of resources available to the resources required to accomplish the task. Since, this algorithm involves reallocation of resources involving VM and data centers. Therefore, every time whenever reallocation occur reauthentication of the resource allocation is also conducted along with reallocation of Load Balancing. So, this algorithm also provides better security, by focusing on the concept of not disclosing the personal details of the user to cloud provider.

4) Anjali D.Meshram et al. [26] proposed that the scheduling strategy should be developed for multiple tasks. In cloud computing processing is done on remote computer hence there are more chances of errors, due to the undetermined latency and loose control over computing node. The author of the paper solves this determines this problem and tries to give a fault tolerant model for the cloud computing. The FTMC model tolerates the faults on the basis of reliability of each computing node. A Computing

node is selected for computation on the basis of its reliability and can be removed, if does not perform well for applications.

The approach used by the author in this paper: Reliability assessment algorithm is applied on each node (virtual machine) one by one. Initially reliability of a node is set to 1. There is an adaptability factor ***n***, which controls the of reliability assessment. The value of *n* is always greater than 0. The algorithm takes input of three factors *RF*, minReliability and maxReliability from configuration file. *RF* is a reliability factor which increases or decreases the reliability of the node. Min Reliability is the minimum reliability level. If a node reaches to this level, it is stopped to perform further operations.

The author concludes the paper with the simulation results showing that the proposed fault tolerant technique works better than other similar fault tolerant techniques.

Some of the strategies proposed by various researchers using the load balancing techniques are:

VectorDot

A. Singh et al. [12] proposed a novel load balancing algorithm called VectorDot. It handles the hierarchical complexity of the data centre and multidimensionality of resource loads across servers, network switches, and storage in an agile data center that has integrated server and storage virtualization technologies. VectorDot uses dot product to distinguish nodes based on the item requirements and helps in removing overloads on servers, switches and storage nodes.

CARTON

R. Stanojevic et al.[13] proposed a mechanism CARTON for cloud control that unifies the use of LB and DRL. LB (Load Balancing) is used to equally distribute the jobs to different servers so that the associated costs can be minimized and DRL (Distributed Rate Limiting) is used to make sure that the resources are distributed in a way to keep a fair resource allocation. DRL also adapts to server capacities for the dynamic workloads so that performance levels at all servers are equal. With very low computation and communication overhead, this algorithm is simple and easy to implement.

Compare and Balance

Y. Zhao et al. [14] addressed the problem of intra-cloud load balancing amongst physical hosts by adaptive live migration of virtual machines. A load balancing model is designed and implemented to reduce virtual machines' migration time by shared storage, to balance load amongst servers according to their processor or IO usage, etc. and to keep virtual machines' zero-downtime in the process. A distributed load balancing algorithm COMPARE AND BAL-ANCE is also proposed that is based on sampling and reaches equilibrium very fast. This algorithm assures that the migration of VMs is always from high-cost physical hosts to low-cost host but assumes that each physical host has enough memory which is a weak assumption.

Event-driven

V. Nae et al. [15] presented an event-driven load balancing algorithm for real-time Massively Multiplayer Online Games (MMOG). This algorithm after receiving capacity events as input, analyzes its components in context of the resources and the global state of the game session, thereby generating the game session load balancing actions. It is capable of scaling up and down a game session on multiple resources according to the variable user load but has occasional QoS breaches.

Scheduling strategy on LB of VM resources

J. Hu et al. [16] proposed a scheduling strategy on load balancing of VM resources that uses historical data and current state of the system. This strategy achieves the best load balancing and reduced dynamic migration by using a genetic algorithm. It helps in resolving the issue of load imbalance and high cost of migration thus achieving better resource utilization.

CLBVM

A. Bhadani et al. [17] proposed a Central Load Balancing Policy for Virtual Machines (CLBVM) that balances the load evenly in a distributed virtual machine/cloud computing environment. This policy improves the overall performance of the system but does not consider the systems that are fault-tolerant.

LBVS

H. Liu et al. [18] proposed a load balancing virtual storage strategy (LBVS) that provides a large scale net data storage model and Storage as a Service model based on Cloud Storage. Storage virtualization is achieved using an architecture that is three-layered and load balancing is achieved using two load balancing modules. It helps in improving the efficiency of concurrent access by using replica balancing further reducing the response time and enhancing the capacity of disaster recovery. This strategy also helps in improving the use rate of storage resource, flexibility and robustness of the system.

Task Scheduling based on LB

Y. Fang et al. [19] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain high resource utilization. It achieves load balancing by first map-ping tasks to virtual machines and then virtual machines to host resources thereby improving the task response time, resource utilization and overall performance of the cloud computing environment.

Biased Random Sampling

M. Randles et al.[20] investigated a distributed and scalable load balancing approach that uses random sampling of the system domain to achieve self-organization thus balancing the load across all nodes of the system. The performance of the system is improved with high and similar population of resources thus resulting in an in-creased throughput by effectively utilizing the increased sys-tem resources. It is degraded with an increase in population diversity.

Active Clustering

M. Randles et al. [20] investigated a self aggregation load balancing technique that is a self-aggregation algorithm to optimize job assignments by connecting similar services using local rewiring. The performance of the system is enhanced with high resources thereby increasing the

throughput by using these resources effectively. It is degraded with an increase in system diversity.

Server-based LB for Internet distributed services

A. M. Nakai et al. [21] proposed a new server-based load balancing policy for web servers which are distributed all over the world. It helps in reducing the service response times by using a protocol that limits the redirection of requests to the closest remote servers without overloading them. A middleware is described to implement this protocol. It also uses a heuristic to help web servers to endure overloads.

Join-Idle-Queue

Y. Lua et al.[22] proposed a Join-Idle-Queue load balancing algorithm for dynamically scalable web services. This algorithm provides large-scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers for the availability of idle processors at each dispatcher and then, assigning jobs to processors to reduce average queue length at each processor. By removing the load balancing work from the critical path of request processing, it effectively reduces the system load, incurs no communication overhead at job arrivals and does not increase actual response time.

4. Comparison of Load Balancing Strategies

In this section there is a comparison among the strategies given by various researchers for balancing the load in cloud computing. The researchers used the load balancing techniques to provide new strategies.

S. No.	Author/Year	Strategy Based on Load BalancingTechnique	Environment	Concept
1	A. Singh et al. / 2008	VectorDot	Datacenters with integrated server and storage virtualization	The node is distinguished based on the item requirement by using dot product.
2	R. Stanojevic et al. / 2009	Carton	Unifying framework for cloud control	The cost is minimized by load balancing and for fair allocation of resources Distributed Rate Limiting is used .
3	Y. Zhao et al. / 2009	COMPARE AND BALANCE	Intra-Cloud	This method is based on sampling and adaptive live migration of virtual machines is used.
4	V. Nae et al. / 2010	Event-driven	Massively Multiplayer Online Games	Complete capacity events are used as input which analyzes its components and generates the game session load balancing actions.
5	J. Hu et al. / 2010	Scheduling strategy on LB of VM resources	Cloud Computing	The best load balancing is achieved and dynamic migration is reduced by Genetic algorithm, historical data and current state of system.
6	A. Bhadani et al. / 2010	CLBVM	Cloud Computing	Uses Global state information is used in making load balancing decisions.
7	H. Liu et al. / 2010	LBVS	Cloud Storage	Replica Load balancing module is achieved which controls the access load balancing and balancing algorithm is used to control data writing load balancing.
8	Y. Fang et al. / 2010	Task Scheduling based on LB	Cloud Computing	A two level task scheduling mechanism is discussed in which load balancing is

				achieved by first mapping tasks to virtual machine and then virtual machines to host resources.
9	M. Randles et al. / 2010	Honeybee Foraging Behavior	Large scale Cloud Systems	Local server action is responsible for Global load balancing.
10	M. Randles et al. / 2010	Biased Random Sampling	Large scale Cloud systems	Random sampling of the system domain is used to balance the load across all nodes of the system.
11	M. Randles et al. / 2010	Active Clustering	Large scale Cloud systems	Self Aggregation algorithm is used to optimize job assignments by connecting similar services by local re-wiring.
12	Z. Zhang et al. / 2010	ACCLB	Open Cloud Com-puting Federation	To achieve better load balancing, small world and scale-free characteristics of complex network is used.
13	S.-C. Wang et al. / 2010	(OLB + LBMM)	Three-level Cloud Computing Network	Opportunistic Load Balancing is used to keep each node busy and Load Balance Min-Min is used for achieving the minimum execution time of each task.
14	H. Mehta et al. / 2011	Decentralized content aware	Distributed computing	The scheduler decides the best node for processing the requests using a unique and special property (USP) of requests and computing nodes. Content information is used to narrow down the search.
15	A. M. Nakai et al. / 2011	Server-based LB for Internet distributed services	Distributed web servers	A middleware is used to support the protocol. This protocol limit redirection rates to avoid remote servers overloading and heuristics is used to tolerate abrupt load changes.
16	Y. Lua et al. / 2011	Join-Idle-Queue	Cloud data centers	Initially idle processors are assigned to dispatchers for the availability of the idle processors at each dispatcher And then assigns jobs to processors to reduce average queue length of jobs at each processor.
17	X. Liu et al. /2011	Lock-free multiprocessing solution for LB	Multi-core	In this method multiple load-balancing processes run in one load balancer.

The comparitive chart presents the concepts used in the strategies for load balancing. By knowing the concept we get to know, that which load balancing technique should be used in various situations.

5. Research Challenges

In this section, we present our analysis of the results obtained through the systematic literature review. Looking at the collected data, we have identified several categories of issues that have been the focus of `research in the past few years.

Security & Privacy
This category includes organizational and technical issues related to keeping cloud ser-vices at an acceptable level of information security and data privacy. This includes ensuring security and privacy of sensitive data held by banks, medical and research facilities. Security and privacy issues become even more serious when governmental institutions use the cloud. Despite the known need for Service Level Agreements between Cloud service providers and users, standards for safety have not yet been established and more research in this area would be beneficial. Security and privacy of data spans issues such as authentication, encryption and detection of malware, side channel attacks and other kinds of attacks —both internal and external to an enterprise. There exists current research on detection and handling of security breaches to guard against tampering, loss and theft of data. Further, fault tolerant mechanisms for backing up data are required when there are failures in the infrastructure, such as network outages.
Other issues, especially in public clouds, include se-cure virtualization through effective firewalls, VM isolation, and detection of reconnaissance scans. There are security issues with long-term storage correctness and migrating from one vendor to another based on changing needs, *i.e*., the problem of vendor lock-in.

Infrastructure
This category entails issues pertaining to the hardware layer used as a backbone for cloud services as well as the thin layer of software used to operate this hardware. The main issue that dominates this category is performance including topics like SaaS placement problems, server allocation optimization, load balancing and many others. Other issues are related to networking such as traffic management, ubiquitous connectivity, network speed and cost, and network reliability. Another group of challenges pertain to resource management including dynamic resource provisioning, scaling, and allocation; as well as resource stranding and fragmentation.
Furthermore, sustainability stands out as another important issue given the amount of energy needed to operate large-scale hardware infrastructure. Quality attributes of the hardware infrastructure have also been an area of interest including issues like availability, reliability, and scalability. Other issues under this category include infrastructure design issues and virtualization.

Data Management
As cloud computing is enabling more data-intensive applications at the extreme scale, the demand is increasing for effective data management systems. One main topic in this category is data storage and all the issues that come with it such as data federation (*i.e*. storage across different providers), data segmentation and recovery, data resiliency, data fragmentation and duplication, and data backup. Other issues include data retrieval and processing, data provenance, data annonymization and data placement (across different data centres).

Interoperability
Most research on issues and challenges with cloud computing recognize interoperability as a major adoption barrier because of the risk of a vendor lock-in. Amongst the many problems being discussed are: the lack of standard interfaces and open APIs, and the lack of open standards for VM formats and service deployment interfaces. These issues result in integration difficulties between services obtained from different cloud providers as well as between cloud resources and internal legacy systems.

Service Management
The Cloud as a service-based IT model created a number of challenges pertaining to service management. Service provisioning seems to be at the core of such challenges. The literature suggests that there is an urgent need for automating service provisioning and making it more dynamic. Automatic combination of services has also been suggested. Another challenge is related to the ability to provide customizable and more context-aware services. The authors in have recognized that managing longer-standing service workflows is a major challenge considering the impact of service failure on numerous complex applications into which the service is integrated. Furthermore, managing service lifecycle and service registry and subscription has proven to be challenging for various reasons.

Virtual Machine Migration
Some vendors have implemented VM migration in the virtualization solution—a big advantage for application uptime in a data centre. What is VM migration? Consider the case of a server with a hypervisor and several VMs, each running an OS and applications. If you need to bring down the server for maintenance (say, adding more memory to the server), you have to shut down the software components and restart them after the maintenance window—significantly affecting application availability. VM migration allows you to move an entire VM (with its contained operating system and applications) from one machine to another and continue operation of the VM on the second machine. This advantage is unique to virtualized environments because you can take down physical servers for maintenance with minimal effect on running applications. You can perform this migration after suspending the VM on the source machine, moving its attendant information to the target machine and starting it on the target machine. To lower the downtime, you can perform this migration while the VM is running (hence the name "live migration") and resuming its operation on the target machine after all the state is migrated. The following are some of the benefits of virtualization in a cloud-computing environment:

- Elasticity and scalability: Firing up and shutting down VMs involves less effort as opposed to bringing servers up or down.

- Workload migration: Through facilities such as live VM migration, you can carry out workload migration with much less effort as compared to workload migration across physical servers at different locations.

Server Consolidation

Server consolidation is an effective approach to maximize resource utilization while minimizing energy consumption in a cloud computing environment. Live VM migration technology is often used to consolidate VMs residing on multiple underutilized servers onto a single server, so that the remaining servers can be set to an energy-saving state. The problem of optimally consolidating servers in a data centre is often formulated as a variant of the vector bin-packing problem, which is an NP-hard optimization problem. Various heuristics have been proposed for this problem. Additionally, dependencies among VMs, such as communication requirements, have also been considered recently.

Economic Challenges

This category includes issues related to the cost-benefit aspect of the cloud from a financial point of view. Some research is focused on producing cost models that reflect more accurately the actual cost of building and operating a cloud. For cloud providers, the cost of the hardware infrastructure and the administrative costs associated with it are key to understanding the economic viability and sustainability of the business. Cloud providers also need to work on effective monetization strategies that would provide a reasonable return on their investments. This includes producing profitable pricing models, resource bundling options and licensing strategies. Moreover, the way billing and payments are currently handled by different cloud providers lacks clarity as to what the customer is paying for in terms of type of service, quality and availability which makes financial benchmarking and comparison across different providers rather difficult.

5. Future Work

There is a future work to solve the load distributing problem on 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. Load balancing 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. Various new algorithms can be proposed for the load balancer so that the load is evenly distributed to every node resulting in better response time and user satisfaction.

6. Conclusion

In this paper we have surveyed the load balancing issue in cloud computing and analysed various techniques used in load balancing. In cloud computing load balancing is the main issue. Load balancing is required to distribute the excess dynamic local workload evenly to the entire node in the whole cloud to achieve a high user satisfaction and resource utilization ratio. It also ensures that every computing resource is distributed efficiently and fairly. There are various researchers who have used the load balancing techniques to propose new strategies. Their work done in the domain of load balancing is analysed and compared. But the issue of load balancing is still open for research work so that high user satisfaction and resource utilization can be achieved.

References

[1] GautamShroff, "Enterprise Cloud Computing," Cambridge University Press, June, 2011.

[2] Qi Zhang, Lu Cheng, RaoufBoutaba, "Cloud Computing: State of Art and Research Challenges," Springer, April 2010.

[3] http://thecloudtutorial.com/related.html.

[4] GolamMoktaderNayeem, Mohammad Jahangir Alam,"Analysis of Different Software Fault Tolerance Techniques", 2006.

[5] L. M. Vaquero, L. Rodero-Merino, J. Caceres and M.Lindner, "A break in the clouds: towards a cloud definition," SIGCOMM Computer Communication Review,vol. 39, pp. 50–55, December 2008.

[6] GeoffroyVallee, KulathepCharoenpornwattana,Christian Engelmann, AnandTikotekar, Stephen L.Scott," A Framework for Proactive Fault Tolerance", Feburary, 2008.

[7] Elvin Sindrilaru,,AlexandruCostan,, ValentinCristea,"Fault Tolerance and Recovery in Grid Workflow Management Systems", 2010 International Conference on Complex, Intelligent and Software Intensive Systems.

[8] M.Armbrust, A.Fox, R. Griffit,et al., "A view of cloud computing", Communications of the ACM, vol. 53, no.4, pp. 50–58, 2010.

[9] Hong Luo, Dejun Mu, Zhiqun Deng, et. Research of Task Scheduling in Grid Computing.Application Research of Computers. 2005; (5): 16-19.

[10] Xindong You, Guiran Chang, Xueyao Deng, et. Grid Task Scheduling Algorithm Based on Merit Function. Computer Science.2006; 33(6).

[11] Wen sheng Yao,et al. Genetic Scheduling on Minimal Processing Ele-mentsinthe Grid. Springer-VerlagHeidelberg. 2002.

[12] Singh A., Korupolu M. and Mohapatra D. (2008) *ACM/IEEE conference on Supercomputing*.

[13] Stanojevic R. and Shorten R. (2009) *IEEE ICC*, 1-6.

[14] Zhao Y. and Huang W. (2009) 5*th International Joint Conference on INC, IMS and IDC*, 170-175.

[15] Nae V., Prodan R. and Fahringer T. (2010) 11*th IEEE/ACM International Conference on Grid Computing (Grid)*, 9-17,2010.

[16] Hu J., Gu J., Sun G. and Zhao T. (2010) 3*rd International Symposium on Parallel Architectures, Algorithms and Programming*, 89-96, 2010.

[17] Bhadani A. and Chaudhary S. (2010) 3*rd Annual ACM Banga-lore Conference*.

[18] Liu H., Liu S., Meng X., Yang C. and Zhang Y. (2010) *Interna-tional Conference on Service Sciences (ICSS)*, 257-262, 2010.

[19] Fang Y., Wang F. and Ge J. (2010) *Lecture Notes in Computer Science*, 6318, 271-277.

[20] Randles M., Lamb D. and Taleb-Bendiab A. (2010) 24*th Inter-national Conference on Advanced Information Networking and Applications Workshops*, 551-556.

[21] Nakai A.M., Madeira E. and Buzato L.E. (2011) 5*th Latin-American Symposium on Dependable Computing*, 156-165.

[22] Lua Y., Xiea Q., Kliotb G., Gellerb A., Larusb J. R. and Green-ber A. (2011) *Int. Journal on Performance evaluation.*

[23] Jasmin James, Dr.BhupendraVerma "EFFICIENT VM LOAD BALANCING ALGORITHM FOR A CLOUD COMPUTING ENVIRONMENT" International Journal on Computer Science and Engineering (IJCSE), September 2012.

[24] JaspreetKaur et al. " Comparison of load balancing algorithms in a Cloud", International Journal of Engineering Research and Applications (IJERA), May-Jun 2012.
[25] NyandeepSran,NaveepKaur et al. "Zero Proof Authentication and Efficient Load Balancing Algorithm for Dynamic Cloud Environment",
International Journal of Advanced Research in Computer Science and Software Engineering ",
7 July 2013.
[26] AnjaliD.MeshramA.S.Sambare et al." Fault Tolerance Model for Reliable Cloud Computing", International Journal on Recent and Innovation Trends in Computing and Communication, July 2013.