CHAPTER – 2

Background and Objective of present work

2.1 BACKGROUND

Today availability is the mail critical point in the field of computing centers. It has been a great challenge to provide the 24 * 7 continuous availability. In past few years lots of research has been done on various systems to provide the alternate solutions for getting the high availability and still lots of researches have been in progress since it’s the critical subject in the field of computing but the goal is same, how to reduce the downtime, cost & increase the availability or continuity of the business applications.

Lots of hardware & software were introduced in the field of computing like hardware load balancers, software load balancers like web servers, clustering which is the group of multiple applications work together with same functionality etc but virtualization is the term specifically designed to explore the capabilities of your hardware resources like the processor, RAM, buses, cpu utilization, cache etc.

Cluster concept was introduced in conjunction with hardware as well with Software. In case of specifically to the hardware, different machines are clustered together to provide the failover as well as load balancing So that in case of failure of any one machine other machines in the cluster will continue to work to provide the high availability, and further the services can be migrated to the other machine in case on interruption of the services in the mid of certain request. But the drawback is same that this setup is very costly because you need a replica of each and every machine to work in conjunction with the other machines. Second you need more labor, more space, more hardware, and more maintenance which is directly in proportion to the cost to organization[5].

Specifically in case of Software, Clustering has been implemented by implementing the parallel logical division of servers called managed servers where the same application distributed equally and work together to provide the failover and load balancing capabilities. Suppose we have a group of four managed servers and we have distributed an application to all managed servers so that in case of failure of any one of the managed server other three will serve the request and in the same way in case of failure of other managed server other servers in the cluster will serve the request. Same concept will be applicable in
the case of failure of a particular instance in between serving the requests. Suppose if a client get connected with any one of the instance in the cluster and in between serving the request that instance got failed then the connection will failover to other instance in the cluster and the complete session with all information will also replicate if session replication enabled.

In the same way hardware load balancer and web servers used to load balance and failover the requests to other load balancer or web server. In case of failure of one of the hardware load balancer the other one will take over the request and in case of failure of one of the web server the other one will take over the request. In such type of architecture the load balancer and web server distribute the requests to internal servers in the round robin manner like suppose if we have four managed servers are in the cluster then load balancer or web server will forward the first request to first managed server, second request to second managed server and so on and after fourth request the fifth one will come to the first managed server in cyclic way[5][6].

Cluster load balancing work on the certain defined manners or you can say algorithms like -

Round Robin

Random

Waeitage based

Round robin as I have explained earlier, worked in cyclic manner, Suppose if we have two servers in the cluster then first request will forward to first managed server and second request will forward to second managed server and the third one will cyclic again to the first managed server. This topology should be taken care at the time of the cluster configuration.

Random doesn’t work on any login, it just forward the request to any one of the managed server in the cluster, suppose if we have two managed servers are in the cluster then the request will forward to any of the managed server, may be first two request on first managed server or one to server one and other to managed server two or both on server two.

Waeitage based work on the basis of the Waeitage defined for each and every server in the cluster. It's useful in case we have some machines with low configuration and some with high configuration, in that case we know a server with good configuration can able to handle large number of requests in
comparison with the server with low configuration. We defined the percentage of load for each and every server during cluster configuration like suppose we have two machines having one first machine with good hardware resources and second one with low hardware resources then we can define the algorithm like 80% of load goes to first machine and only 20% of requests should goes to second machine.

2.2 Objective of Present Work

My idea is to create a new enterprises business application virtual environment to achieve load balancing and failover without investing extra cost or you can say with a non noticeable cost. As I have discussed during introduction that create a new redundant environment for failover is very costly and my approach is based on idea where we can migrate our live running applications from one virtual machine to another in case of the known and unknown outages. Similar to failover in scheduled and non scheduled events, we can achieve proper load balancing by using more than one virtual machines using the high available hardware and software clustered environment. Failover of a virtual machine from one to another would depend on some heartbeats mechanism where the primary and secondary machines will be in synch by sending the heartbeats to each other in a regular interval of 2 or 3 seconds to let other machine know that I am available at the moment and in case primary machine not able to communicate with secondary or secondary found no heart beats in a particular interval then on the based on defined algorithm application will migrate from primary to secondary machine without any outage. In the same way load balancing can be achieved by using t=different virtual machines and by creating redundant application environments. So without any more extra cost we can achieve the failover and application redundant environment for proper failover and load balancing.

Will discuss in details on this under Methodology section.

So business continuity is the main goal of implementing an Disaster recovery and high availability solutions. High availability is really required if your business really needed the availability over 90%, if availability over 90% is not really required then high availability is not really required but since today every organization information technology infrastructure directly related with the revenue so almost each and every organization needed the availability over 90%. So each and every organization has their own criticality and severity depending on their business modules and considering the same then a high availability can be designed and implemented.

It's not really very easy to calculate the exact cost with respect to the outages and a user connectivity to the system is not really meant with the system is down. Cost of the outages depend on lots of facts like for how long the system was down, was the complete functionality was down, was a particular module was down, was there any performance issues and system was responding in the regular interval, or was there any
issues from client or vendor network side etc. A user connectivity with the system always doesn’t mean that system is not available since there could be your systems are running properly but some performance issues from client or vendor network side.

.1) Main objectives of my research work

1. Business application continuity without or with minimum interruption

The primary objective of any highly available business continuity environment is the 24*7 continuity of their business applications without any interruption. Today almost all of the organizations are spending a lot to achieve and build a highly available environment for the continuity of their business applications because it’s directly related with the cost and revenue of the organization.

2. Easy and fast monitoring and fast failures detection

There should be some perfect monitoring in place to detect the failure of the primary machine where your applications are running like I have been working to develop a solution called heartbeats to communicate between primary and secondary failover machine to know each that primary machine is running or not. So in case of communication got lost due to the failure of the primary machine, all of the running applications from primary machine should migrate or failover to the secondary machine without or with very minimum time frame.

3. Auto failover within minimum timeframe

In case of failure detection, the failover from the primary machine to the secondary machine should be in minimum timeframe to make the proper continuity of the business applications. Today the applications took almost 15 to 20 minutes during failover migration and the business applications doesn’t respond in that time frame but yes, if you have implemented a proper cluster highly available solution then you will have zero outage in such kind of failures because other applications in the cluster will respond in case of the failover of a clustered machine.

4. Rapid recoverability in case of failure
Application continuity or recovery should be in minimum timeframe. Application continuity means a user can able to login to the system and should able to perform the regular operations without any issues.

5. **No session data loss with session replication**

There should not be any loss of connected data due to the failure of the virtual machine or application servers. Loss of connected data means, suppose a user accessed an application and doing something on that and mid of a transaction and from the backend the connected machine or application server got crashed in that case the user data should survive and replicate to the another machine or server in the cluster.

6. **Create and implement a best planned operational environment**

During the initial setup of a business environment create a properly highly available architecture with complete failover and load balancing capabilities using the hardware and software highly available cluster services.

Overall you can say the primary objectives of highly available solutions are to provide reliability, maintainability, recoverability, failure or error detection and continuous running business environment.

### 2.3 WORK PLAN

I have divided work plan into many phases in conjunction with the system analysis and design. Initial way is determine the existing and capabilities of the load balancing, failover and virtualization in the current market. The previous key problems, solutions and the current key problems. The reasons why virtualization still has not been implemented by most of the organizations and what we can do further to make better utilization of virtualization and explore the further enhancements.

### 4.1 Key Points

1) Identify and analyze the existing virtualization solutions where it is already deployed and implemented.

2) Identify the things like what are the current virtualization service provide companies are doing in the field of virtualization.
3) Identify the solutions on how the virtualization will improve the failover, load balancing and virtualization.

4) Identify the challenges come in the way to get the trained resources or staffing.

5) How will virtualization will really save hardware, infrastructure, labor and maintenance cost of the IT infrastructure environment.

6) Identify the current issues going on related with the virtualization.

7) Identify how failover can handle in a better way.

8) Identify the support of virtualization with the legacy hardware’s and what we can do further to paired and operate them more efficiently with the virtualization layer.

4.2 Executing work plans

Following are the guidelines will use during the executing of the work plans:

1) Meetings and data collection

The main purpose of the meetings and data collection is to gather the information about the existing implemented solutions, advantages and disadvantages. The nature of current work, if virtualization is currently implemented anywhere or not, if implemented then what are the challenges there, Complete study of the implemented solution, how the load balancing and failover are achieving and the total time frame currently required for the failover with the current implemented solution. Whether the proper load balancing is happening or not and the implemented solution is giving the performance as expected. These all the information can be gathered by interacting with the experts already working in the same environment or in the implementation phase. Once data collected properly then we can analyze the situation properly.

Data collected in the following different way -

- Type of servers (Linux, Solaris, Unix etc.), number of servers and storages
- Topologies and location of the servers, distributed, collocated or consolidated
- Estimation of the hardware and storage utilization
- protection and security levels
2) Understanding and constructing the business cases

Understand the current nature of business, if the high availability is really required or not, if the business needs their application to be available more than 90% then there is a need of highly available solution. So first proper understanding of the business is required and further on the basis of information we can create the cases to analyze further how we can integrate or implement virtualization in the existing environment without any complications.

3) Benefits for the business

Analysis of the complete life cycle for the benefits for business and organization. Will it really going to benefit business from operational as well as from functional point of view.

4) Rules-of-thumb when constructing a storage virtualization business case

Key rules while designing the storage virtualization solution.

5) Key risks and impacts of virtualization implementation

Before implementing virtualization all the involved risks and impacts should be identified properly.

6) Design the solution

A proper design of the architecture.

7) Testing results

Proper testing of the architecture on test environments and in LAB before implementing in the live production environment.

8) Reconciling plans with actual

9) Dissolving the effort