Scaling in Cloud Computing
Rajni Patil, R. K. Singh

ABSTRACT
The cloud computing is a resource and offers computer assets with services instead of a deliverable product which allows storage and sharing of the files of multiple types like audio, video, software’s, data files and many more. The data is shared over internet cloud storage and can be accessed for free and also at an affordable price. The effective way of sharing the information and technology by collaboration the real world to availed the competitive advantages. This paper makes a brief description about the cloud computing and its scaling techniques the main explanation is on vertical scaling and horizontal scaling with examples.

KeyWords
Cloud computing is one of the major technologies in the current and next generation and a huge discussion point. The cloud has become an offering these days attracting the researchers and media on large scale. The fortune corporate have invested huge in the technology of cloud computing for having their own private cloud storage and also public cloud like Google, Amazon and Microsoft.

It is an information technology gateway shaping in new forms. The crisis in the economic strategy in the recent years it has been quite a budget on the companies to invest and continue their storage servers. The introduction of cloud and the investment over cloud seems a big step but considering the situation cloud storage is suggestible storage system. With the

Introduction of cloud computing many small and medium companies have got relief and migrate to cloud storage and get free from the heavy cost of the individual servers.

Kevin & Bill (2012) say that cloud computing is not any deployable application product but is a service that allows wide collection of storage of various formats of files. The cloud computing is majorly divided into three main components as application, connectivity and storage. All the three segments have different roles all together form a cloud which wide spread over internet and is also cost effective. [1, 2]

To improve the performance of cloud computing there are scaling factors. These scaling factors improve the performance and throughputs of the cloud services and accessing. The two main types of scaling techniques are discussed in the thesis they are vertical scaling and horizontal scaling which are also known as scaling up and scaling out. The simulation results explain the server queuing process and power consumption when compared to single server with finite size.

BASIC SYSTEM

Sztrik (2007) explains the queuing system in the single server station is consisting of a queuing buffer of an infinite or a finite size, including an identical server. This basic queuing system is also known as service station and or node some time. The queuing system can be explained as: a server can respond to only one service request at a time by one customer as one instance, in such situations the server is either in ‘busy’ stage or in ‘idle’ stage. In case if all the servers seem busy at the arrival of a new job request then the new job request is kept in buffer, the server has assumption of buffer space availability for new request. The new job request is in waiting state in the buffer for its turn to be allocated the process. Once the executing of the current request is departed one request from the buffer is allocated the resource for its execution. This system follows a queue here and this is a queuing system for request executions. [3]

A primary queuing system can be described depending upon the arrival of the process which requests for the resources. This is characterized with a sequence of time random inter arrivals and denoted by the variables \{A1, A2, ....\}. The inter arrival times are independent and are sequentially distributed and also identical in nature. The function of time inter arrivals can be continues or discrete in distribution. The average time inter arrival are given as: \(E[A]=T_A\). The reciprocal is given as: the average rate of arrival is \(\lambda = 1/T_A\).

The time distribution of inter arrival process is exponential and the process is Poisson in nature. The sequence is given as \{B1, B2, ....\}
which is specified service time of successive jobs. It is assumed that the sequence of the variables is random and function is distributed. The service mean time $E[B]$ could be denoted by $TB$ that reciprocal is given as: service rate $\mu = 1/T_B$.

There are quite many such situations where the process requests do not perform any renewals, the job request arrivals depend on the number of the customers, the job request, and facility of the service. These situations mainly occur during the case of Finite- Source queuing systems for the single server process handling.

$E$- Time Inter arrival, $A$ -Random Inter arrivals, $T$ - Time, $B$- Successive jobs, $\mu$- Rate of Service, $\lambda$ – Rate of Arrival.

There is always a need for new innovation in the computer science area and especially when it comes to cloud computing which is been a swing in this generation of storage. There is a search for improving the performance of cloud storage in both the arena of software and hardware for increasing the throughput. There is demand for throughput gain when the systems are in simultaneous demands. The size of 10 giga-byte pipe line drastically falls down when there is request from the hundreds of systems over the network. The most suggestible solution for increasing the performance and gaining the throughput of the systems and server is by implementing the strategy of scaling. The addition of more hardware recourse is only piling the complications and also is an overload. The scaling is much more suggested in cloud computing environment which is an every expanding storage. The ability for the system scaling is more suggestible under stress when compared to increase of over all throughputs and or aggregate the overall performance of every individual component. The scalability in cloud environment is most commonly handled by either horizontal scaling or vertical scaling behavior.

**HORIZONTAL AND VERTICAL SCALABILITY**

The resource increasing in cloud for restoring and or improving the performance of the application is administrated by either scaling outward which is known as horizontal scaling or by scaling in upward direction which is known as vertical scaling. The scaling factors depend upon the nature and constraints of the resources and requirements. Horizontal scaling is more of additions to machines and the devices on hardware front of a computer system to meet the required demands. Vertical scaling is more of resource addition in virtual computing pool i.e. increasing the virtual CPU, increasing the RAM, and such virtual additions to handle extra demands.

**VERTICAL SCALING (scale up)**
Vertical scaling will be achieved by adding enough resources to the existing physical machinery. The additions will be done for the increment of memory, increasing the storage capacity, adding CPUs and increasing the virtual memory. The increase in hardware resources will make the system run as big box. The hardware upgrade could be done until the limit is reached. The vertical scaling has some physical limitations.

**Fig: Scale Up architecture:**

**Implementation**

Single-server queueing system with fixed run length

Mean interarrival time 1.000 minutes

Mean service time 0.500 minutes

Length of the simulation 1000.000 minutes

Average delay in queue 0.577 minutes

Average number in queue 0.600

Server utilization 0.540

Number of delays completed 1039
HORIZONTAL SCALING

The horizontal scaling is done by increasing the hardware in the machinery. The horizontal scaling is easily achieved by adding new resources as per the requirement and the demands, by adding only the hardware in situations will seldom increase the performance in proportion but there will be increase in TCO (total cost of ownership). When planning on adding new resource there is also requirement of networking tools and additional space for the new resource locations. In horizontal scaling there are quite many chances for under using the additional resources. For example for a particular web application their might be network input output while the server and system memory is at the bottle neck. When there is an addition to the current environment the new addition will remain underutilized. This is a process of adding more nodes to the current system. This is a counter pose to the vertical scaling which is actually an increase in performance by expansion of virtual things and where as horizontal scaling is all about expansion of the real hardware things.

Cloud computing is a massive scalability in horizontal fashion with the available applications taking advantage of the environment. This trend is more about increasing the machinery and refactoring the application which work very well in the horizontal scaling environments which eventually means that the application are enough suitable for the cloud computing. All the application that intends to take advantage of the horizontal scaling has to priority to take care whole application system availability with an assumption of consideration that there are many chances for the individual components to fail.

The example of the load balancing in parallelization is a stateless server with number of web applications accessing the data from the web servers, the incoming workload would be distributed over the pool of the servers connected publicly or privately.
CONCLUSION
The conclusion of the thesis is all about the scaling of cloud services with designing and implementation of the vertical and horizontal scaling when compared to single server with finite size. The performance is vertical computing is double that of basic system with consumption of
50% of power and that of horizontal scaling is 70% consumption of power consumption. The mean queuing in case of vertical scaling is less than the horizontal scaling. When compared between the two i.e. horizontal and vertical scaling it is suggested that vertical scaling is much preferred than the horizontal scaling for the cloud performance improvement.

REFERENCES

