Available at: www.researchpublications.org

Load Balancing with Specialized Server Using Database

Anand A. Chaudhari[#], Mrs.V.M. Deshmukh [*]

[#]*Computer Science & Engineering Department,*
SGBAU University

[1]anand07chaudhari@gmail.com
[2]msvmdeshmukh@rediffmail.com
PRMIT&R Badnera, Amravati – 444701, India

***Abstract*— one of the major problems that have been seen in computer networks is of Network Congestion. Many techniques are available to overcome congestion problem but Time and space complexity is the issue. Time and space complexity increases when the Network Congestion takes place. The proposed work will provide a better solution to the Network Overloading by sorting request based on server type and then if required will be applying the Round Robin Algorithm on a particular system. Concerned to database, efforts have been made to handle request depending upon Read/Write type which can be further brought back to load-balancer server. The best possible IEEE format paper on the above area of interest is tried and implemented which will sought some solution to the overloading problem.**

***Keywords*— Network Congestion, Load Balancing, Round Robin, Computer Clusters, Database**

I. Introduction

In computer networking, **load balancing** is a technique to distribute workload evenly across two or more computers, network links, CPUs, hard drives, or other resources, in order to get optimal resource utilization, maximize throughput, minimize response time, and avoid overload. It is commonly used to mediate internal communications in computer clusters, especially high-availability clusters [3]. If the load is more on a server, then the secondary server takes some load while the other is still processing requests.

Various load balancing algorithms can be implement to manage load in computer networks but most efficient algorithm which proves best in response time and throughput is applied. Here different types of servers will be provided in the clusters and then depending upon the request type main load balancer will handover request to the particular server. Requests are again sorted based on READ/WRITE type to balance load on specialized server [2]. Databases of today have to handle high load while providing high availability. One must ensure scalability and high availability for all components, starting from the edge routers that connect to the Internet, all the way to the database servers in the back end. Load balancers have emerged as a powerful new weapon to solve many of these issues [1], [2].

A computer cluster is a group of linked computers, working together closely so that in many respects they form a single computer. Clusters are usually deployed to improve performance and/or availability over that provided by a single computer. , while typically being much more cost-effective than single computers of comparable speed or availability

II. Load BALANCING

With the advent of the Internet, the network now occupies center stage. As the Internet connects the world and the intranet becomes the operational backbone for business, the IT infrastructure can be thought of as two types of equipment: computers that function as a client and/or a server, and switches/routers that connect the computers. Conceptually, load balancers are the bridge between the servers and the networks [1], [4]. On one hand, load balancers understand many higher-layer protocols, so they can communicate with servers intelligently. On the other, load balancers understand networking protocols, so they can integrate with networks effectively. Server load balancing deals with the load distribution across multiple servers to scale beyond the capacity of one server and to tolerate a server failure.

The balancing service is usually provided by a dedicated program or hardware device (such as a multilayer switch).It is commonly used to mediate internal communications in computer clusters, especially high-availability clusters. Load balancing is the main reason for computer server clustering.

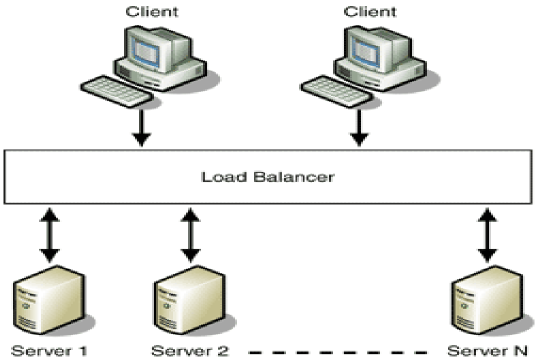

Fig. 1 Block diagram of Load-Balancing [4]

A. *Load Balancing Techniques*

Load Balancing can be categorized into following four types:-

1) *Transport-level load balancing:* Such as the DNS-based approach or TCP/IP level load balancing acts independently of the application payload.

2) *Application-level load balancing:* Uses the application payload to make load balancing decision, It is totally dependent on application [3].

3) *Hardware Load Balancing:* Hardware-based load balancers can route TCP/IP packets to a various in a cluster. These types of the load balancers are often found to provide a robust topology with high availability, but come for a much higher cost.

4) *Software Load Balancing:* Software Load Balancers which can be configured and made more intelligent to accommodate a wider range of load balancing options. Most commonly used load balancers are software based, and often comes as an integrated component of expensive web server and application server software packages [3].

B. *Specialized Server*

There are various types of server available for request handling and hence called as specialized users. Some servers can be listed as:

1) Chat Server: Chat servers enable a large number of users to exchange information in an environment similar to internet newspaper that offers real-time discussion capabilities. Real Time means occurring immediately. For example real time operating systems are systems that respond to input immediately [7].

2) *Fax Server:* A fax server is an ideal solution for organization to reduce incoming and outgoing telephone resources but that need to fax actual documents [7],[8].

3) *FTP Server:* One of the oldest of the internet services, File Transfer Protocol makes it possible to move one or more files securely between computers while providing file security and organization as well as transfer control [8].

4) *Mail Server:* Almost us Ubiquitous and crucial as Web Servers, mail server's move and store over corporate networks via LANs and WANs and cross the internet.

5) Audio/Video (Media) Servers: Audio/video Servers bring multimedia capabilities to Web sites by enabling them to broadcast streaming multimedia content. Streaming technologies are becoming increasing important with the growth of internet because most user do not have fast enough access to download large multimedia files quickly [3],[4].

III. Design & Proposed Work

Specialized Servers arranged in a network will help for easy execution i.e. set-up including main server and cluster computers are made and then depending upon the request type the load is balanced.

Note: Main Server is capable of handling all types of requests. Here we are considering the scenario that:

- Request handling capacity of main server (load balancer) is full i.e. value crosses threshold limit of server and
- Hereafter load balancing will be according to request type. Example Fax Request will be transferred to Fax server and so on.

For Example just consider that we have 4 types of requests as follows: Mail access Request, Media (Audio/Video) Request, Fax Request and Online-Chat Request. These users/clients request will be sorted and forwarded to specific server types without applying any load balancing algorithm. Hence the request handled by the clusters will be look:

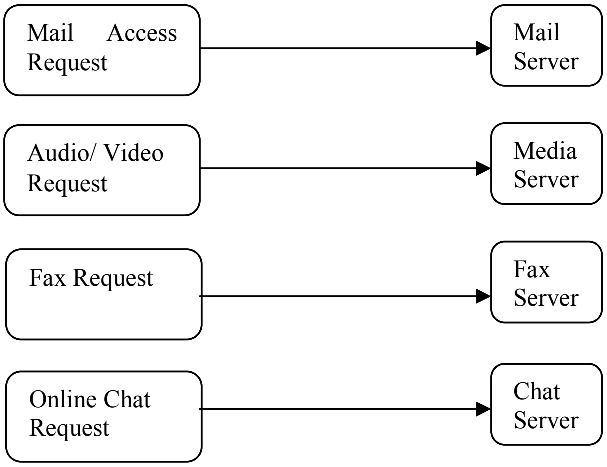

Fig. 2 Load balancing using specialized server

This paper also answers the following question

What if request goes beyond the threshold of these specialized Servers?

We will be having computer clusters to handle overloaded request and that will be according to Round-Robin Algorithm. These cluster computers are normal server which is capable of handling all types of servers.

What is threshold value and how it is calculated?

The load average is the average system load over a period of time. It is conventionally given as three numbers that represent the system load during the last one, five, and fifteen minute periods. Thresholds are baselines established to monitor data collection and application status polling.

Threshold calculation depends upon server machine configuration i.e. CPU handling capacity, H/W configuration etc.

How to categorize incoming client request into Read/write type?

Database plays an important role in sorting request into read and write type and hence again they are forwarded to respective read and write available server.

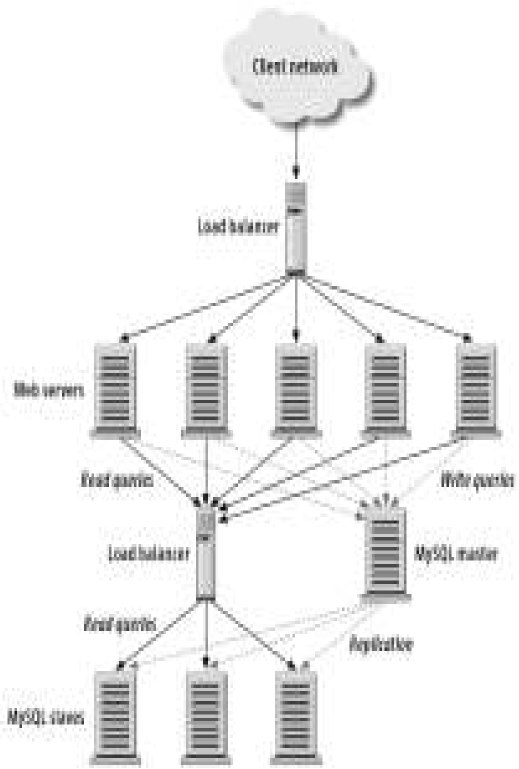

Fig. 3 Load balancing using specialized server.

A. *Round-Robin Algorithm*

Round-robin is by far the simplest algorithm available to distribute load among nodes. It is therefore often the first choice when implementing a simple scheduler. On of the reasons for it being so simple is that the only information needed is a list of nodes [9], [15]. The algorithm traverses the list, returning the nodes one by one. When the end of the list is reached, the algorithm starts from the beginning of the list again [10].

Thus returning the same nodes in the same order. The time-complexity of the selection is O (1). Combined with its low implementation complexity and its low information requirements, the round-robin scheduler is also often the most efficient scheduler algorithm. However this is only when several key assumptions are true [3] – [7]:

1. The nodes must be identical in capacity. Otherwise performance will degrade to the speed of slowest node in the cluster.

2. Two or more client connections must not start at the same time. Should they, the node chosen will be the same, because the order of nodes retrieved from the cluster is the same every time.

3. The jobs must be similar to achieve optimum load distribution among the nodes. If a single node is more loaded than others it will become a bottleneck in the system.

*DNS Round Robin***:** The Domain Name Server (DNS) database maps host names to their IP addresses.

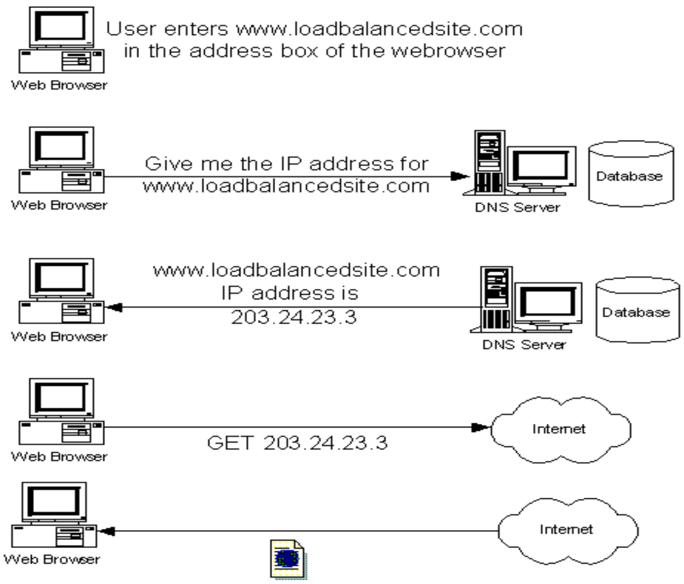
User enters www.loadbalancedsite.com in the address box of the webrowser
Web Browser
Give me the IP address for www.loadbalancedsite.com
DNS Server
Database
Web Browser
www.loadbalancedsite.com IP address is 203.24.23.3
DNS Server
Database
Web Browser
GET 203.24.23.3
Internet
Web Browser
Internet
Web Browser

Fig. 4 Round robin through DNS [6]

Any or all of these assumptions might be challenged in the face of a real application and cluster setup

Round-robin:

+ Fast.

+ Requires only information about the names of the nodes.

+ Does not require any global decisions.

+ Easy to implement.

+ Superb balancing, when the workload is not skewed, nor are the nodes.

− Breaks down completely when the above item does not hold

B. *Computer Cluster*

A computer cluster is a group of linked computers, working together closely so that in many respects they form a single computer. The components of a cluster are commonly, but not always, connected to each other through fast local area networks. Clusters are usually deployed to improve performance and/or availability over that provided by a single computer, while typically being much more cost-effective than single computers of comparable speed or availability [7], [8].

Load-balancing clusters operate by distributing a workload evenly over multiple back end nodes. Typically the cluster will be configured with multiple redundant load-balancing front ends. Since each element in a load-balancing cluster has to offer full service, it can be thought of as an active/active HA (High Availability) cluster, where all available servers process request [3], [4].

Various types of clusters are available and they are listed below:

1) High-availability (HA) clusters:

High-availability clusters (also known as Failover Clusters) are implemented primarily for the purpose of improving the availability of services which the cluster provides. They operate by having redundant nodes, which are then used to provide service when system components fail. HA cluster implementations attempt to use redundancy of cluster components to eliminate single points of failure [8], [9].

2) Load-balancing clusters:

The multiple computers are linked together to share computational workload or function as a single virtual computer. Logically from the user side, they are multiple machines, but function as a single virtual machine [14].

Following are the two key terms associated with Load-balancing:

1) *Latency:*

A performance measure defined here as the non-overlapped portion of network load balancing CPU overhead (lower is better). Latency adds to the client response time [13].

2) *Response Time:*

A Performance measure defined as a round trip delay to process a client request. Response time increases with the non-overlapped portion of CPU overhead, called latency. Amount of time it takes from when a request was submitted until the first response is produced, not output (for- time sharing environment) [13].

IV. Conclusions

In this paper we have proposed a new concept of network load balancing with the help of specialized server. The paper describes how initial sorting of the request is done i.e. basically it is based on request type and then forwarded to specific type server.

In other case round robin algorithm is used for load balancing which proves to be an efficient and best algorithm as far as response time is concerned. Read/Write type client request are sorted and handled using model database. The server load balancing solution provides more advanced and flexible traffic management and stronger processing power as compared to single server implementation.

References

[1] Dennis Haney & Klaus S. Madsen Datalogisk Institut, Kobenhavns Universitet, Fall 2003

[2] Rory Breuk, Gerrie Veerman Database Load Balancing, March 1, 2012.

[3] Tony Bourke, Server Load Balancing O,Reilly & Associates, Inc, August 2001.

[4] Andrew Tanenbaum, Computer Networks (Delhi, Pearson: Dorling Kindersley, 2008).

[5] Chandra Kopparapu, Load balancing Servers, Firewalls and Caches Wiley Computer Publishing 2002.

[6] Ram Prasad Padhy, P Goutam Prasad Rao Load Balancing In Cloud Com

[7] Matthew Syme, Philip Goldie Optimizing Network Performance with Content Switching: Server, Firewall, and Cache Load Balancing July 02, 2003.

[8] Brighten Godfrey, Karthik Lakshminarayanan, Sonesh Surana, Richard Karp, Ion Stoica ”Load Balancing in Dynamic Structured P2P Systems” (C) 2004 IEEE.

[9] David Karger and Matthias Ruhl, “New Algorithms for Load Balancing in Peer-to-Peer Systems,” Tech. Rep. MIT-LCS-TR-911, MIT LCS, July 2000.

[10] Vinod Pisharody, Michael Wu, Tanya Zhiltsova Active-active operation for a cluster of SSL virtual private network (VPN) devices with load distribution.

[11] Anders Olaf Dannie, Michael palmeter Method and Apparatus For Distributing Load on Application Servers.

[12] Mladen Turk: Multilevel Load Balancer.

[13] S. Sharma, S. Singh, and M. Sharma, "Performance Analysis of Load Balancing Algorithms," World Academy of Science, Engineering and Technology, vol. 38, 2008.

[14] [G. R. Andrews, D. P. Dobkin, and P. J. Downey, "Distributed allocation with pools of servers," in Proceedings of the first ACM SIGACT-SIGOPS symposium on Principles of distributed computing Ottawa, Canada: ACM, 1982, pp. 73-83.

[15] S. Malik "Dynamic Load Balancing in a Network of Workstation," 19 November 2000.

[16] N. G. Shivaratri, P. Krueger, and M. Singhal, "Load Distributing for Locally Distributed Systems," Computer, vol. 25, pp. 33-44, 1992.