The Study of Server Load Scheduling Strategy

Junxi Yu
College of Information Engineering
Hebei United University
Tangshan, China
yjxmfc@163.com

Guohuan Lou
College of Information Engineering
Hebei United University
Tangshan, China
zdhua@heut.edu.cn

***Abstract*—In this paper the classification and development of server load balancing technology are briefly described and the load balancing algorithms based on server cluster are compared. A server load balancing technology and algorithm based on multiple parameters are proposed. Finally, the load balancing algorithm is tested. Testing results show that the method is feasible.**

Keywords- server ; load balancing; algorithm; schedule

I. INTRODUCTION

With the continuous development of Internet application and growing of the network information resources, number of users and the network traffic explosively increase. The processing ability and calculating strength of networking core equipment also correspondingly increase, so heavy pressure is brought to a single traditional network device. Relying on hardware upgrade will cause a lot of wasted resources and it is also a big challenge to middle and small enterprises which have limited fund. On the other hand, even though the equipments with more excellent performance also can not meet current business needs. So clusters system with multi-equipment， as its characters of high performance, low price and stronger expansibility, is being used widely.

It has become a pressing issue in a cluster system that how to distribute reasonable business between network devices with same processing capability, how to prevent the emergence of such situation as some devices are too busy but other devices are not in full processing power. So that load balancing mechanism emerges as the times require. This paper mainly discusses load balancing mechanism of server cluster and introduces and compares in detail some algorithms of load balance.

II. CLASSIFICATION OF LOAD BALANCING

The load balancing is based on the existing network structure, to provide a cheap, effective and transparent method of extension of network devices and servers bandwidth, to increase throughput, to enhance network data processing ability, to increase the flexibility and reliability of the network. The load balance of server cluster means the applications of traffic load between all servers and applications in the server cluster. According to devices used, load balancing technologies can be classified into soft load balancing and hardware load balancing. Besides this classification, there are other classification methods. Here are two kinds of classification for cluster load balancing technology [1-2].

A. Classifying by the geographical structure of the application

(1) The local load balancing [2]: The local load balancing refers to making the load balancing for local servers to make use of the existing equipment, to avoid loss of data traffic caused by single server failure. Flexible balancing strategies are used to allocate data traffic to the servers in the server cluster.

(2) Global load balancing [3]: global load balancing refers to the load balancing between server clusters that are placed in a different location and have different network structures. Global load balancing is used primarily in the case that there are own servers in more area , in order to enable global users can access to the server closest to them with only one IP address or domain name. So that the fastest access speed can be obtained.

B. Classifying by TCP/IP protocol layers

(1) Load balancing based on the TCP layer: the distributor monitors a port, when a connection request arrives, depending on the load balancing algorithm the request is transmited to the appropriate server, and the server sends the results to the distributor, then the distributor transmits again to the user. In this way, all users have to access servers through a distributor, so the distributor will become a bottleneck.

(2) Load balancing based on IP layer: this method mainly applies NAT technology (IP addresses conversion). It’s implementation method is very similar to that of TCP layer, that is, monitoring the IP packet. The distributor implements transfer by replacing the source IP address. The distributor is also a bottleneck in this method.

(3) Load balancing method based on DNS: this method is achieved through domain name. When the domain name resolution occurs, according to load balancing algorithms, distributor determines the server that provides services to users and returns to the server address. After the domain name resolution, users connect with the server directly. The workload of distributor is greatly reduced and the work pressure of distributor is alleviated[4] .

Published by Atlantis Press, Paris, France.

III. TRADITIONAL LOAD BALANCING ALGORITHM

A. Round robin algorithm

The requests from the network are assigned sequentially to each server in the cluster, such as server 1, Server 2, Server 3, Server 4, Server 5. As shown in Figure 1.

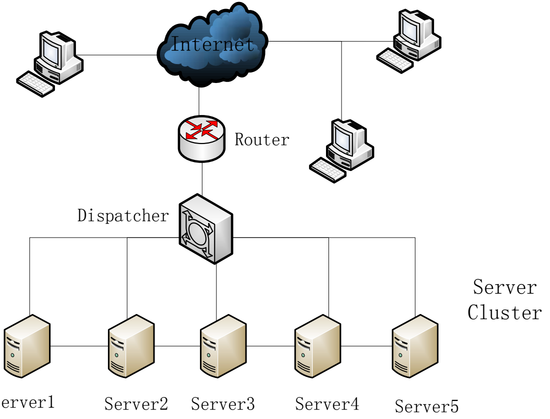

Figure 1. Round robin algorithm

From Figure 1, it can be seen that the advantage of a round robin algorithm is easy to implement. But its shortcoming is also evident. Because the processing capacity of the servers in the cluster is different, this algorithm is absolutely fair with the result that those servers which has stronger processing capacity can not be fully utilized, while the servers with less processing capacity have too many tasks to deal with, even individual server may become incapable. Therefore, round robin algorithm is suitable for the case that processing capacity of each server is similar and the network request is relatively balanced.

B. Weighted round robin algorithm

In this algorithm a weight is added for each server based on the round robin algorithm. This weight represents the processing capability of the server. When assigning task, different number of tasks are allocated to each server according to different weight. For example, the server 1 to server 5, as shown in Figure 2, their weights are respectively 1, 2, 1, 1 and 2.

It can be seen from Figure 2 that the algorithm has improved disadvantages of a round robin algorithm, ensured that the servers with stronger processing capacity can get more tasks, partly to avoid the case that the servers with lower processing capacity are paralyzed due to the accumulation of tasks. However, the algorithm does not consider the time of processing request, and the weight value is given subjectively, which can lead to unbalanced load between servers.

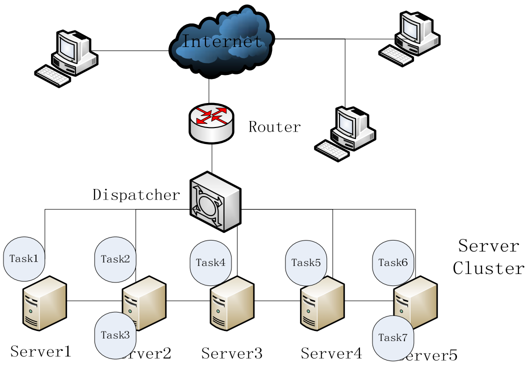

Figure 2. Round robin algorithm with weight

C. Least Connection Scheduling

This method uses a table which records the current task numbers connected by per server, the new service request is assigned to the server that has least connection. As shown in Figure 3, when a new request arrives, according to records of connection numbers in the table, it is assigned to the 4th server.

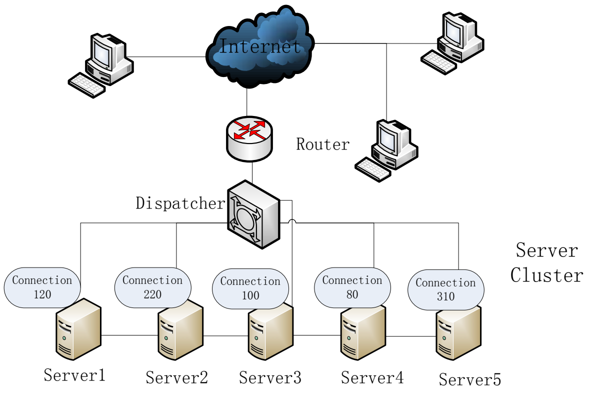

Figure 3. Least connection scheduling

Because of fully taking into account the current connections for each server, least connections algorithm can distribute requests to each server uniformly to achieve load balancing to connections. But it did not take into account the I/O throughput of server in unit time. So thus case may occur that number of connection is large but I/O throughput is small, resulting in server load imbalance.

D. Weighted Least Connection Scheduling

This algorithm is an improvements to the least connection scheduling. Each server uses a weight to indicate its processing ability. While the scheduling is requested, if possible, the task is assigned to the server that has a largest weight and a least number of connection. The weight ratios of server cluster 1,2,3,4,5 are 1:1:2:1:1. The current connecting number and the request scheduling are shown as Figure 4.

Published by Atlantis Press, Paris, France.

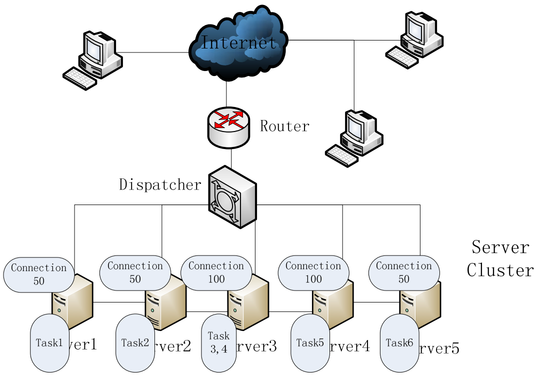

Figure 4. Weighted least connection scheduling

Same as the weighted round robin algorithm, weight of server represents the processing power of a server. But in practice, the setting of weight is related to a lot of factors, only depending on people's experience will bring a big errors.

IV. New load-balancing algorithm and test results

For server load balancing with more parameters all factors which affect server performance should be taken into account. These factors include the server memory consumption when the program is running, current client numbers connected, the number of reading I/O and writing I/O, time-consuming to handle the client request. In addition, the processing delay of each server should be counted (average I/O time). The shorter this time is, the more powerful the server I/O processing ability will be. When scheduling connection requests, the server processing ability should be taken into account according to various factors. If possible, the client service requests are assigned to the server with smaller load.

According to the method above, the experiment is made. Test environment is as follws: a computer with a dual-core Genuine Intel T2080 CPU (1.73GHz), 1024M memory, a computer with a dual-core Intel core i3-380M CPU (2.53GHz), 2048M memory and a computer with a dual-core Intel core i3-2330M CPU(2.2GHz), 2048M memory. The operating system installed in one PC machine is Windows Server 2003. On the other two PC machine the operating systems are Windows XP. The virtual client is installed in one of the PCs, it is used to simulate a large number of clients. One computer is used to simulate the real client and another PC is used to simulate server.

When testing, 100-2000 virtual clients are started respectively, which represent that a different number of clients access the server at the same time. Figure 5 illustrates the experiment results.

During testing the server is at high load environment. The communication program uses IOCP mechanism, First client connects server successfully, the client sends data to ask server providing services for it. After receiving data the server analyses data and then sends the response to client. Next the client sends a request to server again and this process repeats continually. In figure 2 the number of client changes from 100 to 2000 and all of the client and server use this kind of communication way. It can be seen from figure 5 that the characters of server in this high load case are consistent with the expected requirements. As shown in Figure 5, the data processing ability of server varies with the growing clients and the average delay time of I/O increases slowly and linearly , which shows that the server is more stable in I/O delay.

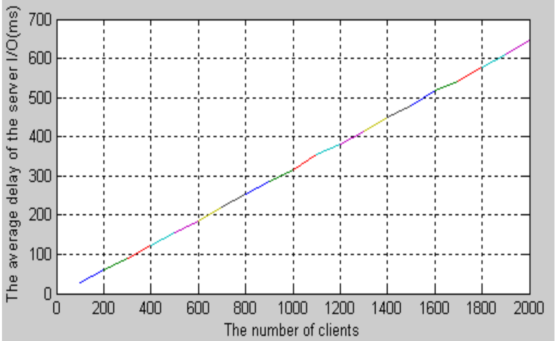

Figure 5. The average delay of server processing I / O request

V. Conclusion

In recent years, the great achievements have been made in the study of load balancing, but with the development of application and architecture, researches need to continue. The load balancing algorithms is still the research focuses at home and abroad, an existing load balancing algorithms must be improved to adapt more complex applications. While designing load balancing algorithms, the different applications should be taken into account. In addition, due to server load has a relationship with a variety of factors, the impact of these factors on the real server applications must be considered.

References

[1] Wei zhen,Wang xue hui,Zhou xia, “Research of content-base distribution in web server cluster,” Computer Engineering and design,2006，27(18). pp. 3410-3412.

[2] Li kun,Wang bai jie. “Research on Load Balancing of Web-server System and Comparison of Algorithms.” Computer and Modernization, 2009,(8), pp.7-9.

[3] Miao yan chao，Zhou ying chao，Hao min. “Video server load scheduling policy.” Microelectronics & Computer, 2004, 21(1), pp.81-82.

[4] Sui pu rui，Jiang jian chun. “Balancing the Load of the Servers in Different Places Bases on DNS.” Computer Engineering，2001，27(3),pp.144-145.

[5] Wu Yongming, He Di. “Design of Bottom Module of Server Based on IOCP.” Information Technology，2007，(3),pp.115-118.