The Designation and Application of a Load Balancing Strategy Used For Session Persistence Based on Dynamic Message Queue

Zhongliang Deng, Xiaoyang Li[a], Fengli Ruan and Wenxu Ma

School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

[a]lixiaoyang0923@126.com

Keywords: Session Persistence, Message Queue, Load Balance, Weighted.

Abstract. In many large-scale enterprise applications, a complex transaction process needs multiple interactions between clients and servers. Because of this, the request from clients should be forwards to the fixed application server, that is, the so-called session persistence. However, to some extent,it goes against the design idea of load balance. For these questions, this paper proposes a load balancing strategy based on dynamic message queue weighted load forecast and balancing distribution model, which can realize input messages balanced distribution effectively according to the load information of servers in clusters and the load state of corresponding message queue while persisting the clients session. Trough theoretical analysis and experimental verification, the results show that this strategy improves the capacity of message parallel processing and client's real-time response.

1. Introduction

With the rapid development of Internet technology, the number of subscribers is increasing, also the network scale is expanding.Recently, the deployment of distributed systems may have been the mainstream solutions for parallel processing of big data and timely response to high concurrent requests.However, the technical difficulties which must be taken into consideration is the designation and optimization of an excellent load balancing strategy according to the application scene.

General Strategy. The dynamic weighted round-robin strategy, based on the least connections or the real-time processing capability and response time of server, calculates the weight according to the real-time loading state of servers.And the input messages will be distributed to servers with lightest loading weight[1].This strategy can realize good load balancing effect, but it can`t persist the transaction process session between clients and servers.Also there is IP Hash load balancing algorithm,which can solve the session persistence problem,through distributing messages according to hash value of clients IP address[2],which can ensure the client request being distributed to the fixed server.But, at a time, there may be problem that a single server with higher load pressure appears.

Aiming to solve the problems mentioned above,this paper proposes a load balancing strategy based on dynamic message queue weighted load forcast and balancing distribution model, which combines the advantages of these strategies mentioned above and realizes input messages balanced distribution effectively according to the load information of servers in clusters and the load state of corresponding message queue while persisting the clients session.Then this strategy will be introduced next.

2. Weighted load forcast and balancing distribution model

Firstly, let`s covert the engineering problem of load balancing into mathematical model, as is shown in Fig.1.Clearly, we can see it is a multiple input and output parallel processing model. $I = \{I1, I2, ..., In\}$ represents multiple input messages. $M = \{M1, M2, ..., M\,\mathrm{r}\}$ represents multiple message queues. $P = \{P1, P2, ..., \mathrm{Pr}\}$ represents the processing model corresponding to message queues.P is mapped to M by server message queue table.The multiple input messages I should be distributed to multiple message queues M,then P reads the messages from the corresponding message queues by searching server message queue table.

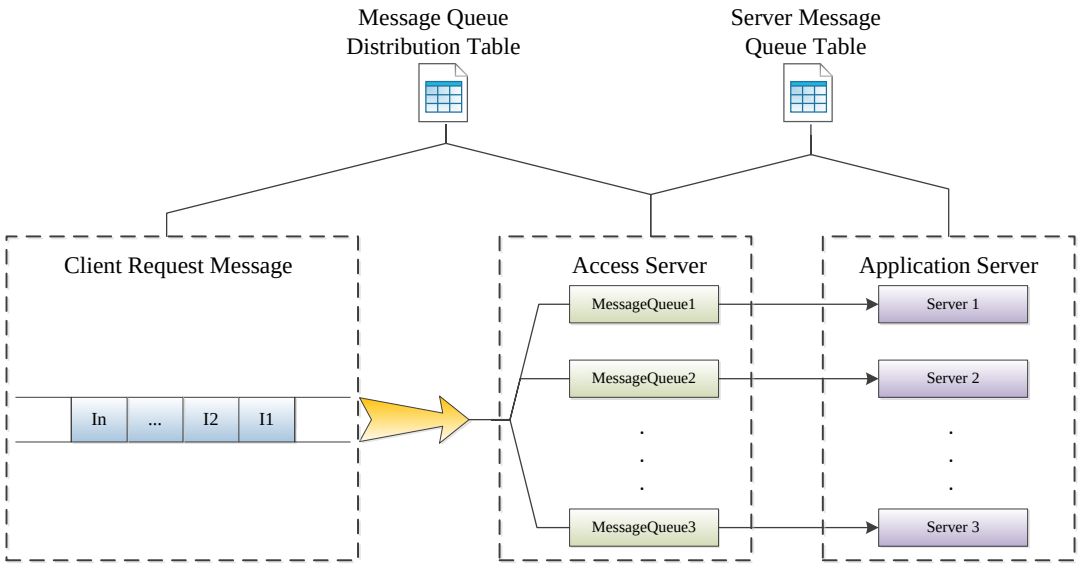

Fig. 1 The mathematical model of weighted load forcast and balancing distribution strategy

The message queue distribution table and the server message queue table ensure the request from client being distributed to the server with lowest load pressure and the session between client and the fixed server will persist until the end of transaction.By this method,it not only guarantees the excellent effect of load balancing,but realizes the session persistence.

Weighted load forcast model. In reference to the general dynamic load balancing strategy, the weighted load forcast model proposed in this paper combines the advantages of both least connections algorithm and dynamic weighted round-robin algorithm.This model calculates the weight of each application server by taking the following four main factors into consideration: 1)the number of connections processed by server,which is represented by N; 2)the usage of CPU,which is represented by C; 3)the usage of memory,which is represented by M; 4)the load state of corresponding message queue,which is represented by F.All of these factors have direct influence to the server[3].Among them,the first three parameters are obtained from the load monitor model in application server.Only the load state F of corresponding message queue is calculated by the weighted load forcast model.Next,we will introduce the calculation process of F.

During the calculation, we have also referred to threshold model.We can calculate the time T that is needed to process the existing messages in the queue by the length of recent message queue and the average time needed to process every message.Then, combining with the inherent forcast algorithm, we can obtain the recent load state value F of the message queue.

We regard last distribution ending time as the beginning and analyze the load state of message queue in the period of $0 \sim \xi$ .Let`s assume thatin this period there is I messages entering into the queue and O messages getting out of the queue and W messages remaining in the queue at ξ time. i_n represents the processing time of NO.n input message. o_m represents thethe processing time of NO.m output message. w_k represents thethe processing time of NO.k message in the queue.From all above,we can learn,in this period,the total processing time of input messages is $T_{in} = \sum_{n=1}^{I} i_n$,the total time of output messages is $T_{out} = \sum_{m=1}^{O} o_m$,and the total processing time of the system in the unit time is

$$T^{'} = (T_{out} - T_{in}) / \xi = (\sum_{m=1}^{O} o_m - \sum_{n=1}^{I} i_n) / \xi . \quad (1)$$

At ξ time, the processing time of messages is $T_\xi = \sum_{k=1}^{W} w_K$ and the average time is $\bar{t} = \frac{T_\xi}{W}$, the length of the queue is $L_\xi = W$.

Accordingto the estimated time mentioned above, we can forcast the load processing time cost of the message queue at any t time.The time cost is $T_t = T_\xi + T'(t-\xi)$ and the length of the message queue is

$$L_t = T_t / \bar{t} = [T_\xi + T'(t-\xi)] / \bar{t} = N[1 + (\sum_{m=1}^{O} o_m - \sum_{n=1}^{I} i_n)(t-\xi) / (\xi \sum_{k=1}^{W} w_K)]. \quad (2)$$

In this model,according to theload processing time cost of the message queue,we can divide the queue into three types of state:the light load,normal and heavy load.Then,according to the threshold model,we define the queue is heavy load state when its length is $L_t \geq \alpha H$,the light load state when its length is $L_t \leq \beta H$ and the normal state when its length is $\beta H \leq L_t \leq \alpha H$ (α is the heavy load coefficient, β is the light load coefficient,H is the capacity message queue)[4].Based on the analysis above,the load state value F of message queue is

$$F(t) = \begin{Bmatrix} 0 & L_t \leq \beta H & \text{Light Load} \\ 1 & \beta M \leq L_t \leq \alpha M & Normal \\ 2 & L_t \geq \alpha M & Heavy\ \text{Load} \end{Bmatrix}. \quad (3)$$

According to the evaluation method of server load state and the importance of the four factors to the load weight, we can deduce the load weight measure formula:

$weight = k^F \times e^C \times N^M$ (K is the load harmonic coefficient and $k \in (1,2)$, $C, M \in (0,1)$). (4)

From the formula, we can also see the influence of the CPU usage to the load weight.

Balancing distribution model.The weight calculated by the weighted load forcast model will be saved into the message queue weight table, according to which, the new client message will be distributed into the lightest load message queue.Fig.2 is the flow chart of the model.

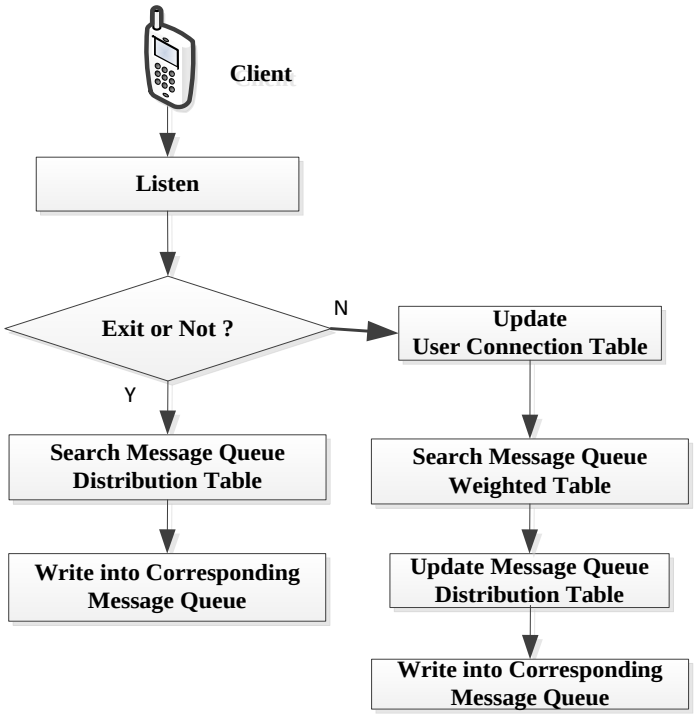

Fig. 2The flow chart of message balancing distribution

Firstly, it will judge whether it`s the first time for the client connecting to the system.If yes, it will update the client connection table and then search the message queue weighted table which is determined by the weighted load forcast model for the minimum weight message queue and then

update the message queue distribution table and write the client message into the corresponding queue.When it connects to the system again,the system will search the message queue distribution table and then write the whole messages from the client into the corresponding queue.The queue and the server is one to one relationship.Then the message queue distribution table ensures the sessions between clients and servers.The new connected client is distributed to the lightest weight message queue according to the message queue weight table,then the purpose of load balancing has also reached.

3. Testing Results.

Now let`s test the effect of the load balancing strategy proposed in this paper.Five message queues are used,which represent the five corresponding application servers in the clusters.Before distributed,the concurrent request messages from clients are saved in the input message queue and the length of which reflects the concurrent requests of clients. Fig. 3 shows the different processing time of each message queue when the length of the input message queue is 10,000.By contrast, clearly we can see, the processing time of messages balancing distributed by the strategy proposed in this paper is obviously shorter and more average than by the dynamic round-robin load balancing strategy or without load balance.

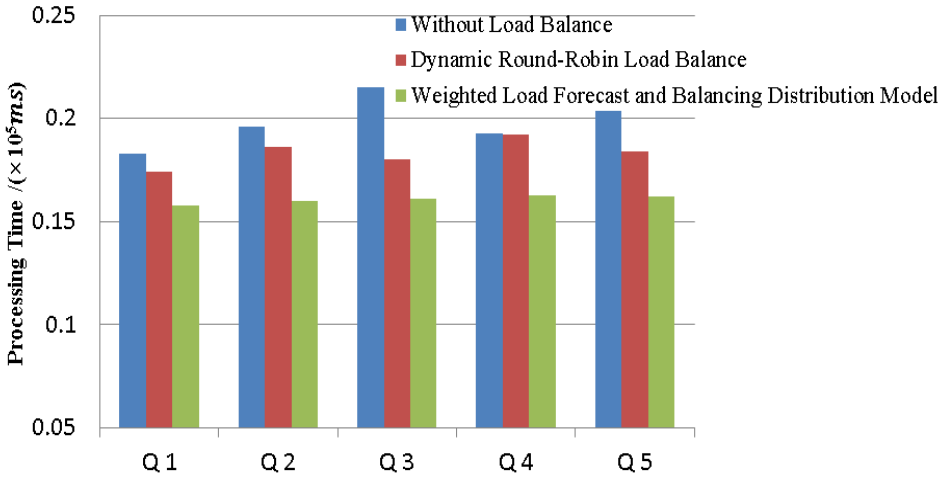

Fig. 3The processing time of each of the five message queueswhen the length of the input message queueis 10,000.

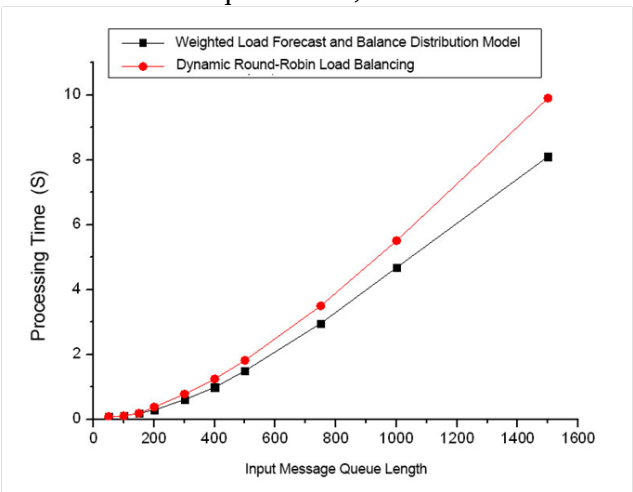

Fig. 4 The comparison of the load balancing effect.

Also we apply this strategy into our project-the designation and implementation of the location based service platform and test the processing response time of the whole system.By giving the different length of input message queue, we have tested the processing time of message queue by

using dynamic round-robin algorithm in general load balancing strategy and the weighted load forcast and balancing distribution model proposed in this paper.The results show in Fig. 4.

As is shown in this picture, when the length of the message queue is small, the processing performance of the two load balancing strategy basically remains the same.However, when the length is greater than 400, the concurrent processing performance of the load balancing strategy proposed in this paper gradually shows out its greater superiority than the dynamic round-robin load balancing strategy.

4. Summary

The paper, firstly, introduces the general load balancing strategy in the practical engineering and points out the problems each strategy may exist.Then it proposes a load balancing strategy used for session persistence based on dynamic message queue, which is better that general load balancing strategies proved by the testing and solves the problem they have encountered. Thisstrategyrealizes the connected clients balancing distributed effectively while persisting the sessions between clients and servers.

Acknowledgements

This work is supported by the National Nature Science Foundation of China (No.61372110) and the National High Technology Research and Development Program of China (863 Program) (No.2012AA120802). The funders had no role in study design, data collection and analysis, decisio n to publish, or preparation of the manuscript.

References

[1] C.Y. Deng, J.T. Zhang, Y.S. Liu, Research on Dynamic Load Balancing Strategy and Corresponding Model,Computer Engineering and Applications. 8 (2011) 131-134.

[2] Y.L.Zhou, F. Liu,Research on Load Balancing of Web-server System, Computer and Digital Engineering. 4 (2010) 11-14.

[3] D.M.Wang, L.D. He, F.F. Liu, N. Su, X. Liu, Message-Oriented Load Balacing Algorithm, Journal of Jilin University(Engineering and Technology Edition). 1(2012) 140-144.

[4] Y. Wang, W.D. Cai, Q. Duan, An Adaptive Dynamic Load Balancing Algorithm, Computer Engineering and Applications. 21(2006) 121-123.