Automatic-Generation and Optimization of Elastic Scaling Rules Based on Neural Network

Wu.DaQin[1, a], Wang.Shanshan[2,b]

[12] *Department of Information Engineering,Jilin Business and Technology College, Changchun, China*

[a] *liulan2003@sina.com,* [b]*15266615@qq.com*

Abstract.

With the cloud computing development, elastic scaling capability is an important factor to ensure the quality of cloud services. In this paper, the author designed resource requirement model about web system based on neural network under the certain quality of service on cloud platforms. According to the model, the method and mechanism for elastic scaling is realized by BP algorithm on cloud platforms.

Keywords: Elasticity; Cloud Computing; BP Algorithm; Neural Network.

0 Introduction

With the rapid development of IT and network technology, cloud computing become a commercial calculation model in recent years, which can provide convenient ways for computing services. Cloud computing is synonymous with pay-for-use pricing structures. Cloud platforms not only ensure the quality of service provided by computing system, but also save idle resources for customers who lack of web resources. So it is necessary to establish automatically elastic scaling mechanism.

1 SLA and Elastic Scaling Mechanism

In order to guarantee the quality of cloud services, cloud providers will sign a commitment—SLA (Service LevelAgreement) with the customers. For example, in the deployment system CPU utilization at each node in should not exceed 95%

and the availability of system no less than 99%. Whether the cloud platforms can guarantee the quality of service provided by computing systems depends largely on capability of elasticity extension. In other words, the cloud platforms are capable to increase computing resources under high-load while release and recycle idle resources under low-load. Elastic scaling can be divided into horizontal scaling and vertical scaling. Horizontal scaling refers to adding or recycling computing resources, while vertical scaling refers to strengthened or weakened calculative ability for the system. In this paper, we discuss horizontal scaling on the level of web server on the cloud platforms.

2 Problem Definition and Form

The key step for realizing the mechanism of elastic scaling is to design the requirement model for resources in the deployment system of web. In order to meet the quality of service specified by SLA, the number of web systems depends on the combination of the characteristic parameter values. As a result, it is influential to analyze and extract the characteristic parameters for the condition of the system running. The resource requirement model can be established according to these parameters, then the elastic scaling rules can be generated.

In web systems, there are usually several web server cluster to handle the requests of users, shown as Fig 1. The load-balancer which is responsible for distributing evenly the requests of the ender users by the specific algorithm works on the layer of server cluster to guarantee the servers running in the same state nearly. For keeping the web server cluster stretch, it is necessary to update the configuration of the load-balance server meanwhile allocate the web resouces.

So far, the process to generate the elastic scaling rules by web systems may come down to:

1) Determine the minimum combination of characteristic parameters, expressed as P1, P2…Pn;

2) Extract historic data sets of elastic scaling operation on the cloud platform in specific patterns, namely the values of P1,P2…Pn, the number of web servers AC, the utility value UV(1 or -1).

3) Analyse the data sets coming from 2) about AC and other components so as to come into the requirement model of web system.

4) According to the model from 3), the elastic scaling rules generated. By the values of characteristic parameters P1,P2…Pn captured in specific condition, after we set UV=1, the numbers of web servers would be obtained.

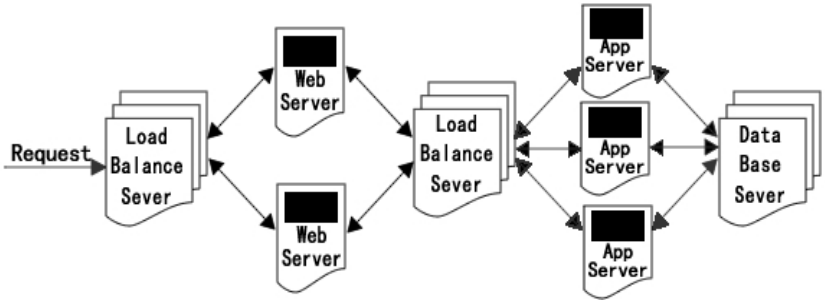

Fig 1 Web System Structure

4 Design of Neural network

In this paper, the model of web resource requirement will be designed by the neural network, then the auto-scaling rules will be received from the model. The relation between the number of web resource requirement and the characteristic parameters can be constructed depending on multilayer feedforward neural network with multiple input and single output.

4.1 Determination of Input and Output layer

The value of input for neural network is the minimum combination of the characteristic parameters, while the value of output is the number of web servers. The ouput layer of neural determines only one neuron.

4.2 Determination the number of hidden layer neurons

To determine the number of hidden layer neurons is mainly based on the complexity of unsolved question appending repeatedly test. Usually, this number is setted about 1.5 times the input dimension. There are 4 input variable in this paper, so the number of hidden layer neurons may be setted 6. It is necessary to fine-tune in light of the actual condition for the best effects.

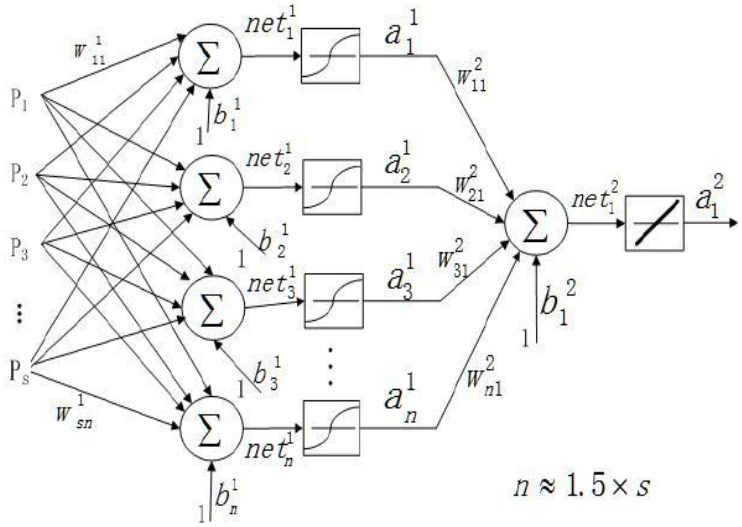

Fig 2 Neural Network Used for Constructing Resource Requrement Model

4.3 Selection Transmit Function and Determination Network Structure

The neural network model is obtained by a function approximation technique, shown as Fig 3. The tangent sigmoid function is selected for the output of hidden layer because the value may be positive or negative, shown as Fig 4. While the linear function is selected for the output layer due to the arbitrary value. Above all, the neural network structure can be designed as shown in fig 2.

The corresponding relationships between variables can be expressed as follows:

$$net_k^1 = b_k^1 + \sum_{i=1}^{s} w_{ik}^1 \times \quad p_i \,, \qquad a_k^1 = \frac{2}{1+e^{-2net_k^1}} \text{-1(k=1…n)} \qquad (1)$$

$$net_1^2 = b_1^2 + \sum_{i=1}^{n} w_{ik}^2 \times a_i^1 \,, \qquad a_1^2 = net_1^2 \qquad (2)$$

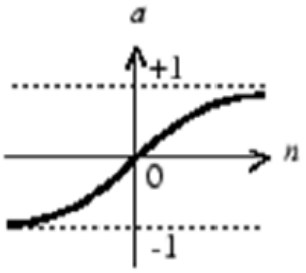

Fig 3 a=tansig(n)= $\frac{2}{1+e^{-2n}}$ -1

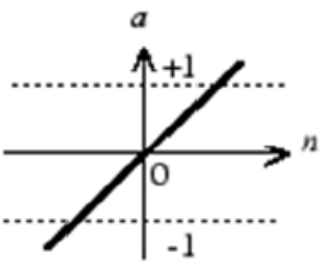

Fig 4 a=purelin(n)=n

4.4 Generation of the Training Sample

In the early stage of generating elastic scaling rules, the sample data from resource allocation need to be accumulated for training the neural network. A simple BP algorithm is used to support the resource allocation, which is executed as follows:

1) Workload-capacity of computing object system, described as Wcap(request/s);

2) Calculate workload change of velocity with time in the past period time, and estimate workload change for some time to come.

3)When the object system need to be allocated resouces, computing the difference between workload from 2) and Wcap, so as to obtain the number of resource allocation and come to the rules for indicating the cloud platform to allocate resource.

4) Record the parameters' values from 3), and estimate the utility value of results(UV) after allocation, then combine the number of resources and them, store to the sample set with the sample formats.

5)When the scale of sample set arrive at SS0, the training process of neural networks is started. Construct the model of requirement resource for generating the elastic scaling rules. Set up reasonable SS0 according to the heuristic experience.

4.5 Optimization Process of Elastic Scaling Rules

It is necessary to adjust and consummate the model for optimizing elastic scaling rules. The training sample sets will be selected for training neural network again. Each training sample set can be described as 1×（n+1）vector. The first n components show the values of characteristic parameter which run in the extensible cloud platform and the utility values of scaling rules which is corresponding with this example. The last one componet shows the number of the web server.

The optimization algorithm of elastic scaling rules can be summarized as follows:

1) The utility value of threshold setting rules UV_0 characterization rules effect is good or bad;

2) Set a variable EC, said until now present continuously extended rules utility values below UV_0, the number of its initial value is zero;

3) After each web system resources allocation according to the elastic scaling rules, UVc will be estimated. If UVc>UV0, EC=0; else EC=EC+1. The information included by rules is combined with discretized UVc to the sample sets in training sample format, then the sample sets should be adjust further.

4) If EC<>EC0, jump 3); else decide if the (EC0+1) time scaling operation has been optimized. If true, give up this optimized process and restart to generate the scaling rules; else start to optimize. Input the training example set to the neural network which is be trained by specific mechanism, then the new model will be obtained when the training is completed. Regenerate elastic scaling rules by the new model which has been optimized. Jump to 2) to continue.

Summary

In this paper, the author designed resource requirement model about web system based on neural network under the certain quality of service on cloud platforms. According to the model, the method and mechanism for elastic scaling is realized by BP algorithm on cloud platforms.

Acknowledgements

This work was financially supported by Jilin Province Educational Science topics " The Research and Realization of Self-help Building Website Platform Based on Saas" (2013399) and "The Tutorial Expert System For Computer Rank Examination Based on AI Technology"(2013542).

References

[1] Armbrust M,Fox A,Griffith R,et al.A view of cloud computing[J].Communications of the ACM,2010,53(4):50-58.

[2] Freitas A L,Parlavantzas N,Pazat J,et al.Cost Reduction through SLA-drivenSelf-Management[C]//9th IEEE ECOWS.Lugano,Switzerland:IEEE,2011:117-124.

[3] Ghanbari H.Exploring Alternative Approaches to Implement an Elasticity Policy[C]//IEEE 4th International Conference on CloudComputing.Washington,DC,USA:IEEE,2011,716-723.

[4] Padhy R P.SLAs in Cloud Systems:The BusinessPerspective[J].International Journal of Computer Science and Technology(IJCST), 2012,3(1):481-488.