ISSN : 0974 - 7435 Volume 8 Issue 8

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(8), 2013 [1130-1134]

A PSO-based resource scheduling strategy for load balancing in cloud computing

Hongwei Zhao
School of Information Engineering, Shenyang University, Shenyang, Liaoning, (CHINA)
E-mail: zhw30@163.com

ABSTRACT

Cloud computing needs to manage a large number of computing resources, while resources scheduling strategy plays a key role in determining the efficiency of cloud computing. It is an important issue how to allocate computing resources reasonably and schedule tasks run effectively which can reduce the complete time and cost of all tasks.this paper built a optimization model for Resources Scheduling of cloud computing system and proposed an improved particle swarm optimization algorithm for the built model. Finally, the result of the experiment indicates that the scheduling system can improve the efficiency of dispatching resource and the utilization ratio in the Cloud Computing system.

KEYWORDS

PSO;
Distribution;
Cloud computing;
Load balance;

INTRODUCTION

Cloud Computing is a computing model, providing resource in resource to users by internet, of which the infrastructure cloud of these resources need not be understood, acknowledged or controlled by users. Also, Cloud Computing is a business model that distributes task to computer resource pool, so that various application systems can obtain necessary computing power, storage space and a variety of software resources accordingly.

From the users' point of view, Cloud Computing system can be divided into two types, i.e., public cloud and private cloud. Public Cloud is operated and maintained by a third party, such as Google, Amazon, and so on, providing resources to users by Internet. Private Cloud, built by enterprises themselves, is generally of small scale, while providing IT resource suitable for business operations. Cloud Computing system put IT resource into package and provide it to users in resource.

At present, a great number of studies and researches have been undertaken in all aspects of Cloud Computing system, such as GFS (Google File System), Hadoop architecture, Amazon's Elastic Cloud Computing system (elastic Cloud Computing, referred to as EC2)and so on, the developing aim of which is to take effective use of geographically distributed resource, while it is very important for optimizing resource utilization to apply to effective resource scheduling strategy.

In the scheduling strategy of Cloud Computing system, we need to coordinately use the distributed resource in which the global agent is responsible to manage local agents, and local agents schedule resource

automatically and transparently. In order to implement the load balancing and improve resource utilization and throughput of the system in the node of Cloud Computing system, how to schedule resource becomes a central mechanism in Cloud Computing system, while the scheduling strategy depends on the load information acquisition and the technology processing computing node information.

Particle swarm optimization (PSO)[1] is an innovative artificial intelligence technique for solving complex optimization problems. This discipline is inspired by the collective behaviors of social animals such as bird flocks. Therefore, how to obtain the load information of each node and how to use PSO measurement and evaluation of the local agent to schedule resource will be the main research directions in clouding system.

RELATED WORKS

This paper focuses on the improvement on the retrieval speed of resource and the construction of Cloud Computing system with high efficiency by studying the scheduling methods and load balancing features under the Cloud Computing environment. Mainly concerning with the research of this paper, related works at home and abroad are presented as below[2,3].

GFS, the abbreviation of Google File System (Google File System) launched by Google, can meet the processing data requirements with rapid growth. Google File System is a scalable distributed file system, suitable for the large-scale, distributed, large amount of data access applications, running on ordinary low-cost hardware devices, while providing resource of higher overall performance and fault tolerance functions with a large number of users. However, a GFS cluster always handles the Master scheduling request by a global agency to, so that the global resource agency will become the system bottleneck under the large-scale Cloud Computing environment with more nodes.

As the first company to provide remote cloud platform resources, Amazon called their cloud system as the Elastic Cloud Computing (elastic Cloud Computing, referred to as EC2). Amazon built its Elastic Cloud Computing on the company's large-scale cluster computing platform, and users can operate the various samples running on the cloud system by Elastic Cloud Computing web interface, by which users control the complete running samples on virtual machines and handle the scheduling requests with the application of distributed approach to of local agents. However, Amazon is lack of Load balance[4] support in the process of scheduling.

Aiming at the dynamic changing features of system scheduling under the resource request of Cloud-computing system, on the basis of analysis and research of load balancing[5-7] scheduling models, a Cloud Computing system scheduling mode of hierarchical loading balance has been proposed, following with the basic architecture of the system form with a detailed design. Finally an efficient resource allocation algorithm has been achieved, which takes not only the number of resource of each local agent but also the performance of the various local agents and the current load into account[8-10].

SCHEDULING SYSTEM OF CLOUD COMPUTING

Particle swarm optimization

The canonical PSO is a population-based technique, similar in some respects to evolutionary algorithms, except that potential solutions (particles) move, rather than evolve, through the search space. The rules (or particle dynamics) that govern this movement are inspired by models of swarming and flocking. Each particle has a position and a velocity, and experiences linear spring-like attractions towards two attractors:

I. Its previous best position. II. Best position of its neighborsIn mathematical terms, the ith particle is represented as $x_i = (x_{i1}, x_{i2}, ...x_{iD})$ in the D-dimensional space, where $x_{id} \in [l_d, u_d], d \in [1, D]$, ld, ud are the lower and upper bounds for the dth dimension, respectively. The rate of velocity for particle i is represented as $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$ is clamped to a maximum velocity Vmax which is specified by the user. In each time step t, the particles are manipulated according to the following equations:

$$v_{id}(t) = \chi(v_{id}(t-1) + R_1 c_1(p_{id} - x_{id}(t-1)) + R_2 c_2(p_{gd} - x_{id}(t-1))) \quad (1)$$

$$x_{id}(t) = x_{id}(t-1) + v_{id}(t) \quad (2)$$

where R1 and R2 are random values between 0 and 1, c1 and c2 are learning rates, which control how far a particle will move in a single iteration, p_{id} is the best

FULL PAPER

position found so far of the ith particle, p_{gd} is the best position of any particles in its neighborhood, and is called constriction factor (Clerc and Kennedy, 2002), given by:

$$\chi = \frac{2}{\left|2-\varphi-\sqrt{\varphi^2-4\varphi}\right|} \quad (3)$$

Where $\varphi = c_1 + c_2, \varphi > 4.$

Kennedy and Eberhart (Kennedy and Eberhart, 1997) proposed a binary PSO in which a particle moves in a state space restricted to zero and one on each dimension, in terms of the changes in probabilities that a bit will be in one state or the other. The velocity formula (1) remains unchanged except that x_{id}, , and are integers in {0, 1} and must be constrained to the interval [0.0, 1.0]. This can be accomplished by introducing a sigmoid function $s(v)$, and the new particle position is calculated using the following rule:

$$\text{if } rand < S(v_{id}), \text{ then } x_{id} = 1; \text{ else } x_{id} = 0; \quad (4)$$

where rand is a random number selected from a uniform distribution in [0.0, 1.0] and the function $s(v)$ is a sigmoid limiting transformation as follows:

$$S(v) = \frac{1}{1+e^{-v}} \quad (5)$$

Layered scheduling system architecture

According to the pros and cons of the above mentioned two kinds of scheduling modes, we have brought out a layered scheduling model, in which the global agent is in charge of collecting load information of the relative local agent and all of the resource is submitted to the global agent, but different from centralized scheduling, not all of these tasks are saved in the global agent resource submitted queue waiting for scheduling, but are directly assigned to local agents by global agent in accordance with load balancing and scheduled by local agents. Thus the global agent will not interfere with the resource and its load reduce, which avoids becoming the bottleneck in the system with its less resource waiting time, in order to achieve a simpler realization than distributed scheduling. From the view of the whole resource Cloud Computing system system, taking centralized scheduling in local parts and the distributed scheduling in global ones would not only maintain the advantages of centralized scheduling, but also make up for the deficiencies of it in the use of distributed scheduling on the overall situation layered scheduling system structure is composed of the following parts:

1) resource distribution module, receiving the requests for resource and distributing it according to various resource nodes in order to achieve the dynamic rational resource distribution.
2) PSO Scheduling, a receiving module to a distributor, in charge of dynamic collection of load information on various nodes, setting up the distribution levels of nodes and transmit the information to resource distribution module on a regular basis by the analysis on the performance of node, node CPU utilization, memory usage and I / O usage, and so on.
3) Monitoring module, monitoring whether the local agent overload or delay too long to start re-scheduling strategy.

Load balancing principle

The main principle of Layered load balancing scheduling is that the global agent would receive all requests for resource, and then distribute them to local agents to implement scheduling based on certain principles, the main purpose of which is to enable the local agent perform a more balanced load distribution in order to obtain a higher overall handling capacity and faster response speed. At present the common methods of request distribution mainly conclude three of them, such as running and turning means, least connection means and fastest connection means. By the first kind of methods, it is simple to achieve but the problem of load balancing has not been put into consideration in essence. By the second one, the performances of various the servers have not been dealt with distinctly. By the third method, the current load condition has not been taken into account though the performance of the server has been considered. So these limitations make these methods fail to achieve the load balancing distribution.

The realization of resource distribution algorithm used layered scheduling system

The resource distribution algorithm used in this scheduling system is mainly composed of three parts, one is to select suitable local agent to distribute resource according to load balancing through global agent after

getting load information on various local agents; the other one is to search resource node base on PSO algorithm ;the last one is to introduce the rescheduling mechanism and through setting the threshold as the border to trigger re-scheduling program as soon as discover delays directly impacting on the performance of Cloud Computing system.

EXPERIMENT AND THE ANALYSIS OF RESULTS

In this paper, the project kit, CloudSim, has been used in the simulation experiment, mainly because CloudSim acting on the simulation test focusing on the scheduling strategy in the Cloud Computing, providing the various basic function components of Cloud Computing and simulating the various basic actions of the function components, which making the developers achieve scheduling simulation with ease by this simulation tool. In the simulation experiment, response time has been compared between running and turning means of the resource distribution algorithm and that mentioned in this paper considering load balancing, the two curves are as follows, (shown as Figure 1)

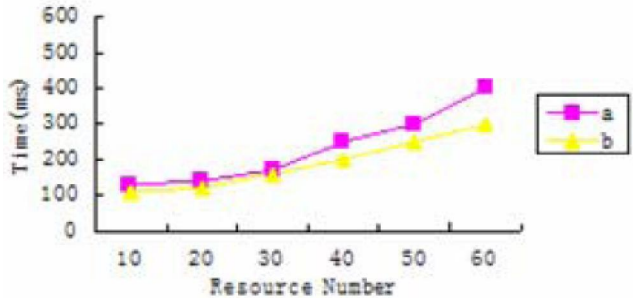

Figure 1 : The comparison chart of two methods

1) a: resource distribution strategy of considering load balancing
2) b: resource distribution method considering load balancing and PSO in this paper

The experimental results show that the Discovery method based on loading and PSO in this paper can effectively reduce the time of processing users' requests in the Cloud Computing system and adapt to the dynamics of Cloud Computing environment by adjusting dynamically the resource discovery method, thereby dramatically improve the performance of Cloud Computing system.

CONCLUSION

In this paper, a kind of dynamic scheduling system supporting Cloud Computing system is proposed in this paper, followed by the analysis of the specific model and technology related to Cloud Computing system, together with a layered scheduling model and the structure of load-balancing system, and then a resource distribution method base on PSO comprehensively considering the task number and current load performance of various local agents has been given out. Additionally, this paper has introduced the re-scheduling mechanism in order to settle the delay problems of resource scheduling in Cloud Computing system through monitoring resource of local agents. Finally, the priority of the method suggested in this paper has been verified by this experiment.

ACKNOWLEDGMENT

The authors were supported financially by the Natural Science Foundation of Liaoning Province (Project No.2013020011) and This work was supported in part by the International S&T Cooperation Program of China (ISTCP) under Grant 2011DFA91810-5 and Program for New Century Excellent Talents in University of Ministry of Education of China under Grant NCET-12-1012.

REFERENCES

[1] D.Karaboga, B.Basturk; On The Performance Of Artificial Bee Colony (ABC) Algorithm. Applied Soft Computing, **8(1)**, 687-697 **(2008)**.

[2] Brian Hayes; Cloud Computing Communications of the ACM., **51**, 9-11 **(2009)**.

[3] Chin H.Yang, A.Dasdan, R.L.Hsiao, D.S.Parker; Map-reduce-merge. Simplified relational data processing on large clusters[C]//International conference on management of data. CA, USA: ACM SIGMOD, **(2007)**.

[4] Ian Foster, Yong Zhao, Ioan Raicu, Shiyong Lu; Cloud Computing and Grid Computing 360-Degree Compared[C]//Grid Computing Environments Workshop. IEEE.12-16 Nov, **(2008)**.

[5] Li.Dongmei, Shi.HaiHu; A hierarchical load balancing scheduling model based on rules computer Sci-

ence, **30(10)**, 16-20 **(2003)**.

[6] Gu.Liming; Research on load balancing technology of Server cluster. Micro-computer information, **4-3**, 20-23 **(2007)**.

[7] M.Clerc; Binary Particle Swarm Optimizers: Toolbox, Derivations, and Mathematical Insights., Available: http://clerc.maurice.free.fr/pso/, **(2005)**.

[8] M.Dorigo, L.M.Gambardella; Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 53-66 **(1997)**.

[9] R.C.Eberchart, J.Kennedy; A new optimizer using particle swarm theory., In: proceeding of the 6th international symposium on Micromachine and Human Science, Nagoya, Japan, 39-43 **(1995)**.

[10] Kang Cheng,Weiming Zheng; Cloud Computing: system instance and Research Journal of Software **(05)**, 1337-1348 **(2009)**.