A Big data dynamic migration strategy

ZHANG Jin Fang[a], WANG Qing Xin[b], DING Jia Man[c*]

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[a]444099725@qq.com, [b]yunnan403@sina.com, [c]tjom2008@163.com

Key words: Cloud computing; big data; load balancing; data migration; internet access; data set

Abstract. In the face of the current resource scheduling using the static load balancing strategy in cloud computing environment could easily lead to the waste of resource, the paper put forward a migration strategy based on the big data migration. The strategy will compare the current load conditions, and choose the data center whose load is lighter, then compute the time consumption of data migration and the number of network access. Finally choose the data which is need to migrate in the date center with lower threshold to migrate the destination node. Experimental shows that the strategy can reduce the consumption time, so as to reduce the proportion of data transmission time accounts for the total execution time.

Introduction

As a new network computing mode, cloud environment is steeply become a new architecture to support big data application [1]. In current definitions, the most representative definition is 3V, that big data should have 3 characteristics: the scale (volume), diversity (variety), and high speed (velocity) [2]. Big data dynamic migration in Cloud computing has catched more and more people's concern [3]. Literature 4 presents a data layout strategy based on clustering matrix, the method using BEA [5] get matrix clustering algorithm, and then for all data sets of collection based on clustering matrix, this strategy can reduce the total number of data transmission across data center in cloud computing environment in the process of execution. 6 of literature reference virtual nodes and a new data distribution strategy based on perception of node capacity has been put forward. In literature 7, it is only for data-insensitive applications in cloud computing. Literature 8 point out a scheduling algorithm that based on time period and budget. Reference 9 put forward a double limit value of the virtual machine migration strategy. Literature 10 shows the efficient data placement strategy and task scheduling strategy based on the correlation of two-phase.

To sum up, the study about big data migration in cloud computing environment is less. Therefore, we propose a network planning and correlation of parameters and fusion algorithm based on probability theory box. resulting in a more rational and efficient metrics related fusion algorithms.

Problem description and modeling

Definition 1. cloud computing environment can be define as: $DC=\{dc_1,dc_2\ldots dc_n\}$. Which is composition by N data centers that is located in different area. This data centers is connection by different network.

Definition 2. dc_i, dc_j and $T(d,dc_i,dc_j)$ respectively: source data center, target data center and time

© 2015. The authors - Published by Atlantis Press

consumption that transmit d from the source data center to target center , the $T(d, dc_i, dc_j)$ can be represented as: $T(d, dc_i, dc_j) = ds / BW\left(link\left(dc_i, dc_j\right)\right) + C_{ij}$ (1)

In the formula, $ds / BW\left(link\left(dc_i, dc_j\right)\right) + C_{ij}$ is the actual time consumption transmit d from dc_i to dc_j. In addition, in the process of data transmission across data centers, there is request , response, connect, disconnect and so on. These can cause some time overhead. To know this part of the time overhead for C_{ij}. Because for the big data application in the cloud computing environment, the data scale is very large. In comparison C_{ij} is very small. So the C_{ij} is ignored that it is represent the time overhead which has nothing to do with the size of the data set. Thus, equation (2) can be used for approximate the time transmit a single data set cross data center one time.

$$T(d, dc_i, dc_j) \approx ds / BW\left(link\left(dc_i, dc_j\right)\right) \quad (2)$$

Data dynamic migration strategy

The process of data migration scheme can be divided into the following three steps:

(1) According to the load of data center, get target data center whose load is low;

(2) Judge the time consumption transmit data sets from the source data center to the target and ordered target data centers by increasing time consumption;

(3) Define target number range according to the amount of data that needs to be moved and the current network access number, and then get the target data centers according to the Threshold .

Phase 1:Record the load information of each data center, including load capacity, the actual load and so on. Target data center: to cut 20% of the workload in the lightest data center as the target data center, notes for DCout.

Phase 2:calculate time consumption of transmit the need to be migrated data sets from source data center to each target data center, then ordered target data centers by increasing time consumption.

Phase 3:Record the network access number as InternetVisit in the process of migration. So:

$$InternetVisit \le \sum_i^n IV_i / T \quad (5)$$

In the $\sum_i^n IV_i / T$, T is for the total time as so far . $\sum_i^n IV_i$ said that divide the total time into equal n, record the number of network access in cloud computing environment as IV_i, than sum it. So we can determine the target centers number N: $N \le \sum_i^n IV_i / T$ (6)

Collect the front N data center in DC_{out} as DC_{out2} whose time consumption is less (increasing ordered by time consumption). The number of the target data centers increase or decrease will cause the change of time consumption and load. Therefore, we need to get a balance between these three

goals.

$$Threshold = a\,Time\cos t + b\ fz + g\ InternetVisit \quad (7)$$

Algorithm 1:

$input: DC_{out2}$

$output: Threshold\{_{[]}\}$

$1.if(i = i < 1, i_{out2}dc++, i)$

$//i$ is for choose the front n。

$2.compute\ Threshold\{i\}$

$3.minThreshold\{\ \} = Threshold\{i\}$

$4.if\ Threshold\{i+1\} < Threshold\{i\}$

$5.minThreshold\{\ \} = Threshold\{i\}$

$6.endif$

$7.print\ f\ Threshold\{\ \}$

$//$ the number in the threshold is the final migration strategy

The simulation results

The experiment was done on the CloudSim simulation platform. Create ten data centers (data center configuration is as follows: CPU of virtual machine is 4 nuclear, host memory is 16G). Before perform a task, these data sets has been distributed in the six data centers based on Zipf distribution. In the experiment, the parameter is the changed quota. All the experimental data in the figure is the average of 20 times experiment.

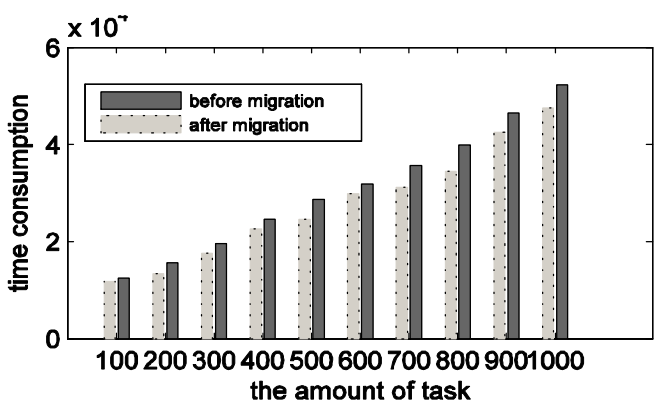

Figure 1:completion Time

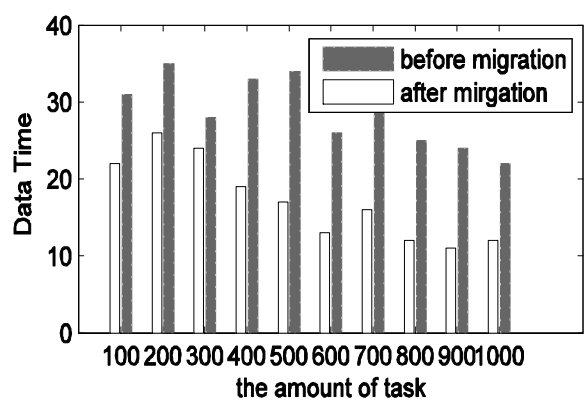

Figure 2 :proportion

In figure 2，the Time of proposed scheme is less than the original.

In figure 3，to complete the task use the proposed migration plan, the data transmission time accounts for the proportion of the total execution time is reduced.

Conclusion

This paper verify the proposed big data migration strategy through extension the CoudSim cloud

computing simulation platform ,Solved the problem of the source waste caused by some resources are idle and the problem of imbalance caused by some data center is heavy load in cloud environment. By compared the proposed strategy and the original distribution, the strategy improved the utilization of resources and ensure the balance of the load.

Acknowledgements

This work was financially supported by the National Science Foundation of China (51467007), and the Application Basic Research Plan in Yunnan Province of China (2013FZ020). The corresponding author of this paper is Ding Jiaman.

Reference

[1] Armbrust M, Fox A, Griffith R, Joseph A, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M. A view of cloud computing. Communications of the ACM 2010 ; 53 (4): 50 –58.

[2] R.Buyya,C.5.Yeo,5.Venugopal. Market-oriented Cloud Computing: Vision, Hype and Reality for Delivering IT Services as Computing Utilities[A].The 10th IEEE International Conference on High Performance Computing and Communications[C],September 25-27, 2008:5-13, Dalian, China.

[3] Dumitrescu CL, Foster I. GangSim: A simulator for grid scheduling studies. Proceedings of the IEEE International Symposium on Cluster Computing and the Grid, Cardiff, U.K., 2005; 1151–1158.

[4] Yuan D, Yang Y , Liu X, Chen J J .A data placement strategy in scientific cloud workflows .Future Generation Computer Systems, 2010,26(8):1200-1214.

[5] McCormick W T , Sehweitzer P J , Whit e T W .Problem decomposition and data reorganization by a clustering technique. Operations Research , 1972, 20(5):993-1009.

[6] Zhou Xiao-li,Zhou Zheng-da.improved data distributed strategy for cloud storage system ［J］.journal of computer applications,2012,32(2):309-312.

[7] Zheng Pai,Cui Li-Zhen,Wang Hai-Yang.A Data Placement Strategy for Data-Intensive Applications in Cloud.[J].Chinese Journal of Computers.2010,33(8).

[8] Liu Ya-qiu，Xing Le-le，Jing Wei-peng.Schedule algorithm based on deadline and budget under cloud computing enviroment[J].Computer Engineering 2013,39(6).

[9] Fang Yi-qiu，Ge Dao-hong,Ge Jun-wei.Research on schedule stratrgy based on dynamic migration of virtual machines in cloud enviroment[J].Microelectronics and Computer,2012,29 (4):45-48.

[10] Liu Shao-wei,Kong Ling-mei,Ren Kai-jun,Song Jun-qiang,Deng Ke-feng,Leng Hong-ze.A two-step data placement and task scheduling strategy for optimizing scientific workflow performance on cloud computing platform[J].Chinese Journal of Computers,2011, 34(11):2121-2130.