Load balancing method to reduce the operating temperature for the reliability of the server

Ryosuke Yamato (1210056)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 12, 2014

Keywords: Reliability, Load balancing method, Virtualization, Cloud computing.

1 Introduction

In recent years, computing services, such as grid computing and cloud computing, have attracted attention. One of the features of the computing services is that the CPU utilization of servers in the computing services frequently becomes 100%. However, the reliability of the servers decreases because it relates to temperature and the CPU utilization.

Many studies reduce power consumption by decreasing the number of running servers. However, the CPU utilization continue to operate at 100%. It does not take into account the reduction of reliability. It does not calculate the failure time of the server.

In this thesis, a method is proposed to improves the reliability of servers. As the number of running servers increase, the operating temperature become lower. On the other hand, the power consumption of servers increase. Therefore, the proposed method aggregates virtual machines into the servers to reduce power consumption while satisfying the reliability required by user. Because the proposed method is aimed at application to the cloud computing, the load of the proposed method is a per-virtual machine. In other words, this thesis solves the optimal placement problem of virtual machine. If the maintenance of the servers is difficult or the

cloud computing provider focuses on the continuation of the service, the proposed method is effective.

2 Proposed method

This thesis introduces a Low Operating Temperature (LOT) load balancing method to improve the reliability of servers. The LOT load balancing method aggregates virtual machines into the servers to reduce power consumption, which relates to the reliability. The policy of the aggregate is that the operating temperature does not exceed the threshold.

System model consists of multiple server groups and a network management server. The network management server is connected to the multiple server groups. The network management server manages virtual machine placement and monitors the temperature of the server in the multiple server groups. The network management server explores a deployable server for a virtual machine.

The LOT load balancing method assigns a virtual machine a server in the multiple server groups when CPU utilization and operating temperature of the server do not exceed 100% and the threshold , respectively.

3 Evalution

This research focuses on aluminum electrolytic capacitors. Because the aluminum electrolytic capacitors are vulnerable to heat, the life of the aluminum electrolytic capacitors is shorter than that of other equipment. According to calculation with aluminum electrolytic capacitors life prediction calculation tool, failure time of 0.01% reduces with the increase in surface temperature and ambient temperature. Based on a graph generated from the calculation with the prediction calculation tool, the constraint condition of the operating temperature is determined.

I evaluated the LOT load balancing method by comparing with a simplified Yajima method. The simplified Yajima method integrates virtual machines into the server without considering the reliability. the simplified Yajima method assigns a virtual machine to a server in the multiple server

groups when CPU utilization of the server do not exceed 100%. The operating temperature model which is created based on the actual measurement data and estimation program are created for the evaluation. I created operating temperature model based on the actual measurement data. The operating temperature model is different depending on the equipment.

In the estimation model, there is one server rack (1U × 40). The maximum number of running servers. The virtual machine request are known in advance. In the estimation, it does not take into account the memory and the network bandwidth and image size. The no running servers shifts to the standby mode.

Firstly the LOT load balancing method assumes, the failure time of 0.01 % assumes set as 10000h. Then the exhaust temperature threshold is 45.0 ℃ from the graph of the failure time of 0.01% and the operating temperature. After that, estimation program is executed. From the estimation results, the number of running servers and the highest exhaust temperature are 30 units and 44.8 ℃, respectively. Final result is that the failure time of 0.01% is 10368h from 44.8 ℃ which is maximum exhaust temperature. On the other hand, results of the simplified Yajima method is that the number of servers running is 27 unit, the highest exhaust temperature is 46.0 ℃. The failure time of 0.01% is 9026h.

In summary of the LOT load balancing method, although the number of running servers increases, but the operating temperature become lower. Therefore, failure time of 0.01 % increases. The LOT load balancing method satisfies the reliability required by the user.

After that, I evaluated the performance of the LOT load balancing method is evaluated with CloudSim. Here, this simulation generates multiple virtual machines. memory and bandwidth and image size are different. This simulation assigns virtual machines to servers with the LOT load balancing method and the simplified Yajima method.

From the results, The LOT load balancing method satisfy the reliability required by the user under all conditions. However, if there is a constraint strong physical resource shortage in addition to the operating temperature threshold, the number of virtual machines in a server decreases. In this case, LOT load balancing method is invalid. In order to effectively use LOT load balancing method, it is necessary to ensure sufficient physical

resources.

4 Conclusion and future work

This thesis has introduced a method, called Low Operating Temperature (LOT) load balancing method. The LOT load balancing method aggregates virtual machines into the servers improve the reliability of servers. The policy of the aggregate is that the operating temperature does not exceed the threshold. As the estimation results of the LOT load balancing method, the reliability of servers improved. because the operating temperature became low by increasing the number of running servers. The number of running servers increase, but the operating temperature is low. The reliability is improved. A cloud environment simulator CloudSim is used to evaluate the performance of the LOT load balancing method. From the results, this research clarified applied scope of LOT load balancing method. Also, the air conditioning power consumption and the failure time of equipment are calculated with CloudSim.

Since this thesis examined only aluminum electrolytic capacitor, it is necessary to also examine other electronic components. Because this research did not create the complex thermal model, it is necessary to create the fluid model. Further, it is necessary to simulate the dynamic virtual machine.