Sun Grid Engine (SGE) and its application

Yinqiao Yang[1,a]   Yanfeng Chen [2,b]

[1]Gansu Normal University for Nationalities, Hezuo, gansu, 747000, China

[2]Gansu Normal University for Nationalities, Hezuo, gansu, 747000, China

[a]Yueguangli7@163.com, [b] Yueguangli7@sohu.com

Abstract

This paper first analyzes the constitution and function of sun grid engine(SGE) art job management system and then analyzes the network structure and the communication time is an important factor affecting the performance of a cluster. Based on the environment of cluster server, put forward the art job migration model and restoration algorithm of cluster nodes, the concrete realization of the system has been obtained in the store a lot of pictures of the art art job management system.

Keywords: Sun Grid Engine; Art job management system; Art job migration; Node recovery

0. Introduction

Cluster computing service object is many applications or work, so assignment management becomes an important part of the cluster [1]. Job management relates to every stage of the art job from the submitted to complete results, related to the various in terms of clusters, it is one of the research cluster technology. Cluster system can use ordinary PC as nodes, cost low and achieve very high speed of operation, improve the system scalability and availability, meet the growing demands of information service.

1. Cluster art job management system

1.1 Introduction of cluster art job management system

Art job management system is a kind of system software based on the operating system, its main role is to strengthen the art job management function of operating system, provide a new mechanism of art job submission, scheduling, execution and control, in order to more efficient use of system resources, balancing network load, improve the overall performance of the system.

1.2 The architecture of SGE

SGE is a grid framework system [2], it is developed by SUN company. The system can be effectively managed by the load on the cluster environment work, the realization of controlling the shared resources, so as to complete a variety of target of enterprise, such as powerful computing ability, efficiency etc. SGE manages resources and strategies, on the one hand the system resource utilization and throughput rate of the system to achieve the maximization, on the other hand it can support the function of the deadline of art job, the priority of art job, user shares resource according to the proportion. The environment of SGE is composed of various shared resources, it uses UNIX series operating system[3].

The SGE software consists of the following components: Qmaster, schedd, execd, shepherd, dcommd, shadowd, Qmon, Qsub[4], each component is a process, station in the corresponding nodes in the cluster and the some core components will be as a background process to carry on the long-term unremitting service. Figure 1 is a schematic diagram of the SGE components.

1.3 SGE operation process

A art job execution process is as follows:

(1) The execd of each node reports to Qmaster about the load information of respective nodes.

(2) The user submits a art job to the Qmaster through the Qsub component.

(3) Qmaster sends the load information and the emergence of new operating system to schedd.

(4) Through the art job scheduling strategy, Schedd uses all aspects of system information, the receive operation is mapped to a suitable node. This will generate the command table back to Qmaster.

(5) Qmaster transmits the art job to the execd of destination node specified by schedd.

(6) The execd create a shepherd for the art job, through the shepherd process is responsible for the management and control of the art job execution.

(7) When the operation finishes, execd will report the art job execution situation to Qmaster.

(8) Qmaster records the resource usage of art job in the database.

(9) shadowd monitor Qmaster, prevent the latter collapse.

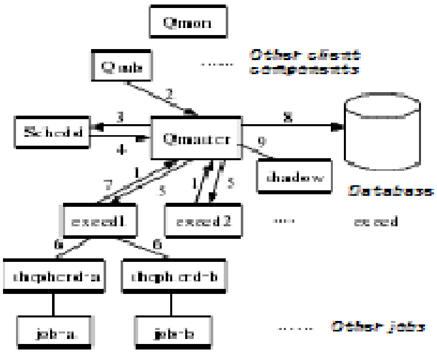

Fig.1 The system structure of SGE

2. The realization of the art job management system based on Cluster Technology

Application of cluster technology in art job management system, it is necessary to solve two problems: one is when a node of cluster system fails, the art job how migration to normal node; another is the recovery process after the failure of the cluster nodes.

2.1The migration process of art job in a cluster system

Job *A* and job *B* are the jobs of the client submitted to a different cluster nodes, in the implementation process, if the node of job *A* fails, then the job actively migrates to the node of the priority is higher (as shown in Fig. 2).

(1) In the normal state, the different network client each submitted job A and B, they run in the respective nodes.

(2) If the node 1, because the system outage, operating system error or application error and other reasons, result in node 1 can't finish the work.

(3) Node 2 gets error information, report to the cluster manager, at the same time, the cluster system software transfers customer jobs to node 2 to run.

(4)When node 1 returns to normal, the cluster system software will return the job of running on node 2 to node 1, job *A* runs on the machine.

2.2 The recovery process of the cluster nodes

When a cluster node occurs failure, the cluster nodes are able to change from failure to recover process; at the same time, the job migration to the new node and run on a new node [5]. The system uses the following procedure completes node recovery process.

(1) Regular receive the events of cluster node from the cluster monitor.

(2) If the cluster events begin on other machines, return (3); if the cluster events are the cluster node failure, return (5); if the cluster events are node recovery, return (9); if the cluster events are the cluster nodes stop, return (14).

(3)Check the shared disk, if normal, return (4), or return (14).

(4)Start running environment of transfer art jobs on the cluster nodes, and performs the migration art jobs, return (1).

(5)Check the shared disk, if normal, return (6), or return (14).

(6)Reset the cluster nodes.

(7) Job migration to the highest priority node; and set the cluster events begin on other machines.

(8) Call processing procedure, and set the cluster events for the node recovery, return (1).

(9)Check the shared disk, if normal, return (10), or return (14).

(10)Reset the cluster nodes.

(11) Job return to the original node; start running environment, access to resources, begin operation.

(12) Call processing procedure and set the cluster events for cluster node stop, return (1).

(13) Processing cluster node SHUTDOWN, the operation environment of migration art job stop, recycling the occupied resources; return (1).

(14) Processing program of shared disk, the cluster server stop.

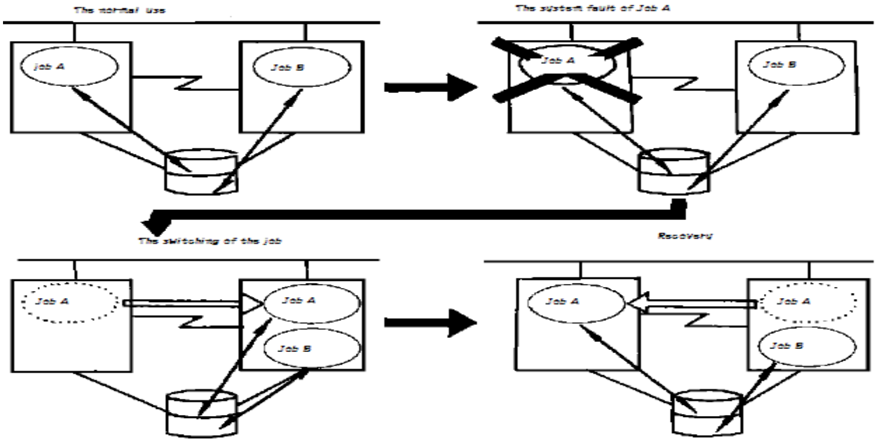

Fig.2 The job migration process of cluster system

3. The analysis of simulation

Parallel computing environment uses mpich-1.2.7. Install Server, Schedd and mpich-1.2.7 on the server. In the test, the number of nodes is obtained from 1 to 7, test ten times, and calculate the average value and the speedup ratio. The principle of mpich-1.2.7 with CPI program is the integral formula $\int {}^{1}\!/\!{\left(1+x^{2}\right)} d_{x} = \pi / 4$, so it can be calculated by integration function $f(x) = {}^{4}\!/\!{\left(1+x^{2}\right)}$ to achieve the purpose of calculating π value. Hypothesis 0 to 1 interval is divided into *n* equal parts, calculate the area of a number of small rectangular of the transverse coordinate between 0 to 1and the longitudinal coordinate between0 to 4. The larger the value, the calculated approximation is more accurate. The test results are shown in table 1. Figure 3 shows the relationship between the ratio of acceleration and the number of nodes, when *n* is larger, the time is used to calculate is far greater than the communication time, the acceleration of parallel computing is linear growth. However, with the increase of the number of nodes, greatly increased for the communication time between each node, when the number of nodes increases to a certain extent, the speedup growth is not obvious. Therefore, for large scale operation, because the proportion of the serial computing part is reduced, use high performance computer to calculate acceleration ratio is larger, the efficient of solution is higher.

Table1The test results

the number of nodes	1	2	3	4	5	6	7
Tq (*n*=10000)	0.00501	0.00579	0.00628	0.01081	0.01061	0.01212	0.01521
Tq (*n*=100000)	0.0346	0.0213	0.0143	0.0175	0.0173	0.0155	0.0171
Tq (*n*=1000000)	0.3178	0.1627	0.1113	0.0904	0.079	0.0658	0.0582
Tq (*n*=10000000)	32.348	16.111	10.741	8.154	6.451	5.396	4.52

Tq (*n*=100000000)	323.888	163.675	119.014	82.165	67.015	53.476	45.449
S(*n*=10000)	1	0.862	0.811	0.439	0.457	0.431	0.315
S(*n*=100000)	1	1.648	2.163	1.809	2.054	2.165	1.85
S(*n*=1000000)	1	1.95	2.896	3.491	4.315	5.009	5.35
S(*n*=10000000)	1	2.011	3.021	3.975	5.02	5.996	7.016
S(*n*=100000000)	1	1.998	2.981	3.875	4.944	5.974	6.994

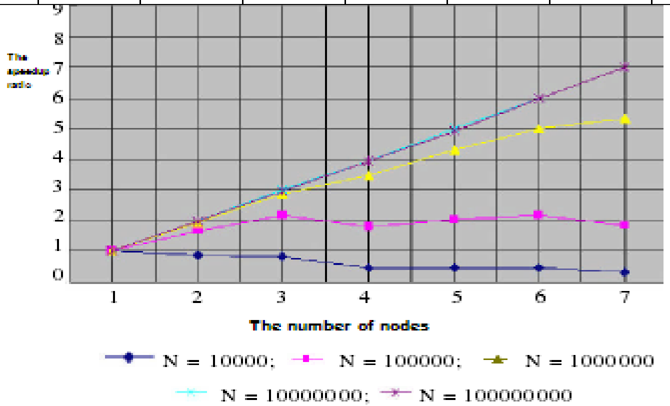

Fig.3 Parallel speedup ratio

4 Conclusions

The SGE job management system achieves the functions of batch job management, and it can monitor submitted jobs well, the utilization of the system has been greatly improved.

References

[1]Victor H L. Cluster Computing: A Survey and Tutorial. California: Miller Freeman Inc, 1997, 19(3), 138–153.

[2]Zhang Chuanfu, Liu Yunsheng, Zhang Tong. Study on simulation grid and scheduling based on SGE [J]. Computer simulation. 2006, 23 (6): 274-278.

[3] Liu Jianfeng. The analysis and research of communication system in grid computing[D]. Harbin: Harbin Institute of Technology, 2002:18-26.

[4] Samad T. High-confidence control: Ensuring reliability in high performance real-time systems, Intelligent Systems[C]. Proceedings First International IEEE Symposium, 2002.

[5] Lu Kezhong, Lin Xiaohui. The implement method of load balancing in MPI parallel programming[J]. Microcomputer information, 2007, 23 (5): 226-227.