2012 International Conference on Image, Vision and Computing (ICIVC 2012)
IPCSIT vol. 50 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V50.51

A Remote Hot Standby System of Oracle

Qiu Long-jin+ and Gao Guang-qiang

College of Computer Science, Sichuan University, Chengdu, China

Abstract. For the problem of database hot standby system of oracle based on shared storage is difficult to ensure disaster recovery capacity, this paper proposed a remote host standby system of oracle. The system has features of remote hot backup, disaster recovery. It ensures the service continuity and disaster recovery capability of the server of oracle database. Experiments show that this system has high stability and capabilities of disaster survivability.

Key words: hot standby system; disaste rtolerant solution; remote backup; service continuity;

1. Introduction

The security and disaster recovery capability of oracle database are very important aspects for enterprise applications [1], many applications currently using hot standby system to ensure continuity of service. The mainstream oracle hot standby scheme is based on shared storage and two servers in the same local area network generally. When the master server can not provide services due to failure, it can switch to standby server quickly. The apparent drawback is that such scheme does not have disaster recovery capability, once the disaster occurs and the storage devices of oracle server are damaged, the data will not be recovered.

The remote oracle hot standby system, referred to ORDB, is a network-based remote backup and recovery system [2]. Data changes of the backup disk on oracle master server (OMS) can be capture in real-time, then packaged into the Backup-Record and sent to remote control center (RCC) through the internet. RCC storage thread will write the back up records cached in buffer queue to the storage logical volume according to the *taskid*.The synchronization thread will send Backup-Records to the oracle standby server (OSS) to archive the data synchronization, storage or synchronization thread will plus 1 to the Backup-Record's Using-Count after the completion of a manipulation of the Backup-Record, the buffer queue will remove the Backup-Record while it's Using-Count equal to 2.

This system can recovery the oracle database even if the OMS and the OSS data corruption at the same time, having a high capabilities for disaster recovery. Meanwhile, the real-time synchronization of OMS and OSS ensures the service connectivity .It is inexpensive but reliable and safe.

2. System Architecture

2.1. Architecture

OSDB is consisted of oracle master server (OMS), oracle standby server (OSS) and remote control center (RCC) .The architecture is shown in Figure 1.

Corresponding author. Tel.: +13666272102.
E-mail address: jinlong2204@yahoo.cn.

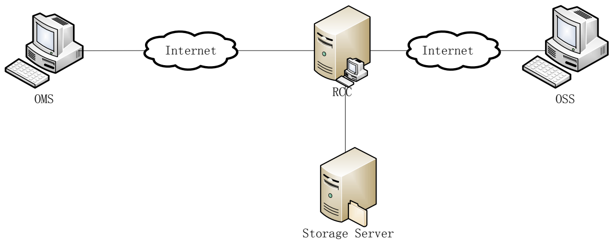

Fig. 1 OSDB system structure

The responsibility of OMS is handing normal business of the database. We installed OSDB client on the OMS for the task of synchronization, recovery and monitoring data changes on local disk volume. OSS needs to install oracle with the same configuration as OMS, even the storage path of database files. The service of OSS is stopped when the service of OMS is normal. After installed the same version of the OSDB client, the RCC will send backup data of the OMS to the OSS constantly to achieve real-time synchronization of two servers in different places. When OMS be in fault, start oracle services and database instance on OSS, then set the IP of the database server to *OSSIP*, OSS will become the main server to provide services, and OMS will become standby server after data recovery by the control center. RCC maintain the task information, create storage logical volume corresponding *taskid* and control the synchronization, recovery, services switch, thus ensuring the service connectivity and capabilities of disaster survivability.

2.2. Modules

OSDB client is consisted of disk management module, security control module and network communication modules. The remote control center is consisted of security control module, network communication module, storage management module. The module association is shown in Figure 2.

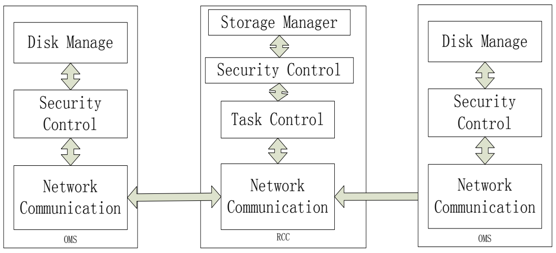

Fig. 2 Functional Block Diagram

The client's disk management module is made up of the massive cache and volume filter driver. Massive cache [3] is a memory-mapped file and can cache Backup-Records that needed to synchronize to the remote control center. Using memory mapped file as a large cache is mainly to improve the efficiency of the system IO. Volume filter driver[4] monitor the OMS data files, log files, control files and configuration files on the disk partition by intercepted writing IRP package, extract the data and package into Backup-Record which will be stored in massive cache and then sent to RCC use communication module to achieve real-time synchronization. When conducting data recovery, disk management module is responsible for extract the valid date from Backup-Records and write them into the disk volume according to the disk offset.

Safety control module checks the network transmission errors of Backup-Record to ensure the accuracy of data transmission. Security control module computes the MD5 checksum of Backup-Records using MD5 algorithm, the Backup-Record format is:

```
Struct record{
BYTE taskid;
DWORD offset;
BYTE * data;
DWORD len;
MD5 h;}
```

taskid indicated the task id, a pair of OMS and OSS can only belong to a *taskid*, and RCC can control multiple tasks; *offset* corresponds to offset of valid data in the OMS disk partition or OSS storage logical volume; *data* is a pointer which point to valid data; *len* is the size of record; *h* record the MD5 checksum of the Backup-Record, MD5 checksum calculation formula as:

$$h = \mathrm{MD}5(taskid, offset, data, len) \tag{1}$$

When the RCC or the OSS receives the Backup-Records, MD5 checksum calculated using the formula 1 is compared with h in the Backup-Records, if the checking passed the data in the Backup-Records is to be written to the appropriate location according to the offset, or request a retransmission.

RCC Mission Control module is used to control the creation of new task, modify task configuration, initializing synchronization, real-time synchronization and recovery, the main components of the module is APACHE, PHP ,MYSQL and WEB management system, when users access the system needs to install a ActiveX control provided by the system which used to send control message to OMS, OSS or RCC, the corresponding control message service thread parses the message received and does the corresponding operation, the task structure as follows:

```
    Struct task{
    PTSTR taskid;
    TCHAR vol;    //disk
volume
    PTSTR OMSIP;
    PTSTR RCCIP;
    PTSTR OSSIP;   }
```

RCC storage management module get OMS partition size when the new task is creating, the storage server then create the storage logical volume with the same size for storing backup data, which will be recycled when the task is deleting. Storage server can be disk array, NAS and other storage technology, this paper uses the disk array [5] as a physical storage server.

3. EY Processes for Achieving ORDB

3.1. Initializing Synchrony

When creating a new task, you need to configure the parameters of structure task in the WEB management system, RCC create the corresponding storage logical volume, then the Initializing synchrony Must be started, which get the data of local disk volume where the database files storage then capture it into Backup-Record, OMS client network communication module will send it to RCC, RCC will write the Backup-Records to the storage logical volume and sent it to the OSS, thus the RCC and the OSS will have the same database files as OMS. The OMS client synchronization algorithm pseudo code is described as follows:

```
Procedure synchro:
SET totalLength=GetDiskLength(task.vol)
SET blockSize=64KB //the length of
backup block
SET offset=0; //current volume offset
While(offset<totalLength)
{  record  r=new record;
   r.taskid=task.taskid
      r.data=GetRecord(task.vol,offset)
   r.offset=offset
   r.h=MD5(r.taskid,r.offset,r.data)
   sendQueue.push(r)
}
While(!sendQueue.isempt())
{ record r=sendQueue.pop();
  Send r TO RCC;
}
```

RCC put the received Backup-Record into receive buffer queue, then the Backup-Records will be processed by storage threads and synchronization thread, storage thread find the corresponding logical volume by the *taskid* and write data to it according to the *offset*, synchronize thread forward Backup-Records to OSS, OSS will check the value of MD5 after receiving the record, if the checking passed, the valid data in Backup-Records will be write to local disk volume according task ‘s *vol* and *offset*.

3.2. Real-Time Synchrony

Real-time synchrony [6] is the core of the system, real-time synchrony guarantees the consistence of database file between OMS, RCC and OSS real-time, making the remote hot standby achieved. the core module of real-time synchrony is the volume filter driver, the driver is located in block device layer, intercepted all packages which type are IRP_MJ_WRITE IRP, extract valid data and package into the Backup-Records, OMS real-time synchrony client flow chart shown in Figure 3 instructions.

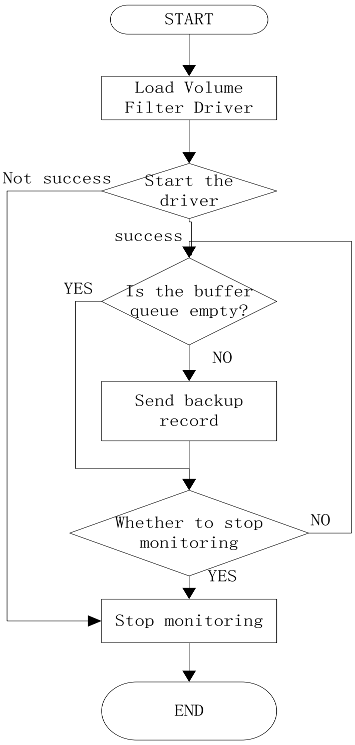

Fig. 3 Real-time synchrony flow chart

3.3. Service Switch

When the primary server failure or completely damaged by the disaster, switching to the standby server is the main purpose of a two-machine standby system. But the two-machine hot standby system is not disaster tolerant, if the storage device has been devastating damage, we can not recover data and services in the short term. This is the main issues addressed by ORDB, after failure of OMS, we can switch to the OSS to achieve service continuity immediately, does not require any data recovery. Service switch depends on the ORDB failure detection algorithm;

OMS periodically send a status packet *S* to the RCC, the structure of *S* as follow:

```
Struct S{
TIME  ts //send time
TIME  tr   //receive time }
TYPE  sta    //statue  of  OMS
service
```

sta represent the statue of oracle service on OMS,which can be USABLE or STOPED.

RCC defined a time window *T* for each task, the size of *T* is 100 times of the state packet transmission period, the collection of state package received *S* is stored in the time window, defined as follows:

$$S = \bigcup_{i=1}^{n} s_i \quad n = [1,100] \tag{2}$$

RCC start a judgment once every 10 cycles for whether to run failure detection algorithm, the determine conditions as follows:

$$\sum_{i=1}^{n} t_i \le T/2 \quad t_i \in \{T' \mid s_i \ne null\} \tag{3}$$

If the formula (3) hold, then start the failure detection algorithm, first calculate the average transmission time difference *tv*, calculated as follows:

$$tv = \sum_{i=1}^{100} (s_i.ts - s_i.tr)/100 \quad s_i \in S \tag{4}$$

If the average time span *tv* is more than 10 cycles, then we concede network failure and continue to wait, or empty the time window T, run the failure detection after 100 cycles, formula

$$\forall t_i \quad \exists s_i \ne null \wedge s_i.sta = USABLE \quad s_i \in S, t_i \in T \tag{5}$$

If Equation 5 is not met, we determine the OMS service is not available, so start the OSS database service and database instance, set the oracle server IP to OSSIP, the standby server becomes the primary server, origin server after the resumption of a standby server.

3.4. Data Recovery

When the OMS, OSS damage occurs, you need to start the data recovery. First, RCC finds the corresponding logical volume by *taskid*, the data block in logic volume is packaged into Backup-Record and sent to the OMS or OSS, after completion of data file recovery we can start the service of oracle, if there are some inconsistencies generally, because all the files needed to recover database, so we can resolve the inconsistencies by oracle’s mechanisms, such as rollback or media recovery.

4. Experiment

To test ORDB time required for service switching, the author deployed the experimental environment of OMS, OSS and RCC, they are not in the same LAN, with independent IP, we install the oracle 10g and ORDB client on OMS and OSS, the ORDB server and web manage system on RCC, the operation system of OMS and OSS is Windows Server 2003, and Linux Enterprise 5 on RCC. Experimental steps are as follows:

1) Install oracle on the D drive In the OMS, built the database instance TESTDB, stored configuration files, control files, log files and data files in the E drive, create the test table, and then install the client, stop the all services of oracle;
2) Install oracle on OSS with the same configuration as OMS, but do not build the table;
3) Built management environment for RCC control center, logon WEB management system, then establish a new task on the OMS, OSS's E drives, set the member of struct task such as *OMSIP*, *OSSIP* and other information, and then delete all the files in the E drive on OSS, start full synchronization of the OMS, when the E disk data synchronized to the RCC the OSS and RCC will have complete copies of the database file on OMS ;
4) Start data monitoring on the E disk in OMS for real time synchronization, then run the stored procedure to simulation DML, DDL operations on database, some times after make OMS power-down;
5) Because of RCC detected that OMS has not sent status packets in 100 cycles, so determine OMS failure，then send the start command to OSS, OSS start the service of oracle.
6) OSS send a message to inform RCC that standby database startup completed, RCC set the primary database server IP to OSSIP, the connection access to the OMS will be switch to OSS, then start the recovery of the OMS.

The author carried out 10 repeated experiments, recorded time required to complete the switching from the OMS power-down to the OSS can provide oracle service, and the data is shown in Table 1.

Table 1 Times needed to switch services

counts	time(s)
1	759
2	541

3	499
4	587
5	647
6	698
7	867
8	602
9	751
10	549

Shown in Table 1, ORDB system can be completed in about 10 minutes service switch, the system can not only prove ORDB complete database synchronization, data backup and disaster recovery, but also a very good warranty service connectivity.

5. Conclutions

We have presented an architecture for remote hot standby system of oracle, which can backup all the oracle database file needed to recover a database on the remote server in real-time. When the Oracle Master Server collapsed, we can switch the oracle service to the Oracle Standby Server in about 10 minutes. Because we have three copies of the database file, even the OMS and OSS are damaged at the same time, we can also recover the database using the backup data in RCC. ORDB integrated the advantage of the two-machine hot standby system and the disaster recovers system, ensure the disaster tolerance and services continuity of oracle services.

6. References

[1] Velpuri, Rama. Oracle backup & recovery handbook [M], 1997.

[2] LAWLER C M, SZYGENDA S A, THORNTON M A. Techniques for disaster tolerant information technology systems. Proceedings of the 1st Annual 2007 IEEE Systems Conference[C]. Honolulu, HI, United States, 2007. 333-338.

[3] GE Liang, LU Zheng-tian, YI Gu-wu, ZHOU Yu. Mass buffer in net work backup system [J]. Applicati on Research of Computers, China, 2009, 26(1):9-12.

[4] WANG Y L, LI Z H, LIN W. A fast disaster recovery mechanism for volume replication systems [A]. The 3rd International Conference on High Performance Computing and Communications[C]. Houston, TX, United States, 2007. 732-743.

[5] Peter M. Chen , Edward K. Lee , Garth A. Gibson , Randy H. Katz , David A. Patterson, RAID: high-performance, reliable secondary storage, ACM Computing Surveys (CSUR), v.26 n.2, p.145-185, June 1994 .

[6] Lloyd S J, Joan P, Jian L, et al. RORIB: An Economic and Efficient Solution for Real-time Online Remote Information Backup[J]. Journal of Database Management, 2003, 14(3): 56-73.