Recovery from storage failures

In our current RAC architecture storage failures are the most problematic ones. In some cases they can lead even to cluster downtime.

The most common reasons of storage failures/unavailability are:

- storage's controller failure
- storage's power suppliers failure
- FC cable failure
- FC switch failure

Although in 2 last cases the storage itself is working, it is not visible from cluster node so from the RAC point of view it is broken.

Note: In most cases storage failures result in a need of its replacement with a spare one. However, as such replacement is quite time consuming and troublesome one should always check whether there is a possibility to fix the problem by other means. E.g. for Infortrend storages sysadmin team has a bunch of spare parts so it is always worthy to check whether the storage can be repaired.

Recovery from a failure of a storage containing cluster registry and voting disks

This is the most malicious failure which results in cluster downtime. Cluster registry and voting disks are vital for RAC operation and they cannot be mirrored on ASM level as the rest of the data, so any unavailability of the storage containing them stops the cluster.

To speed up recovery from such failures, during cluster installation, we create and leave unused on the second storage used by the cluster, partitions that can be used to recreate registry and voting disks.

Recovery time from such situations can be further improved by performing backups of voting disk with 'dd' command as it is described in the note 279793.1 on Metalink. Backups of cluster registry are taken by Oracle automatically every 4 hours. Automation of voting disk backups taking is currently under investigation.

Recovery procedure if a backup of the voting disk is not available

1. Remove CRS installation

a. as oracle on the first node of the cluster copy OCR backup from the default location to a safe place:

```
$ mkdir ~/ocr_backups
$ cp /ORA/dbs01/oracle/product/10.1.0/crs/cdata//* ~/ocr_backups
```

b. As root on all cluster nodes execute the following commands:

```
$ /ORA/dbs01/oracle/product/10.1.0/crs/install/rootdelete.sh
$ /ORA/dbs01/oracle/product/10.1.0/crs/install/rootdeinstall.sh
$ rm -Rf /etc/oracle/scls_scr
$ rm -f /etc/inittab.crs

$ ps -ef | grep crs
$ ps -ef | grep evm
$ ps -ef | grep css
```

If any of the processes exist - kill them.

```
$ rm -rf /var/tmp/.oracle
$ rm -f /etc/oracle/ocr.loc
$ rm -rf /ORA/dbs01/oracle/product/10.1.0/crs/
```

c. As oracle on both nodes use OUI to de-install CRS:

```
$/ORA/dbs01/oracle/product/10.1.0/rdbms/oui/bin/runInstaller
```

2. Check what is the OS name of the LUN containing partitions dedicated for registry and voting disks:

```
$ sudo /sbin/fdisk -l
```

Usually its `/dev/sdb` or `/dev/sdj`.

3. To be sure that partitions discovered in the step 2 are empty clear them with 'dd' command (as root from one node only) e.g:

```
$ dd if=/dev/zero of=/dev/sdb1 bs=1024 count=100000
$ dd if=/dev/zero of=/dev/sdb2 bs=1024 count=10000
$ dd if=/dev/zero of=/dev/sdb3 bs=1024 count=10000
```

4. If necessary change mappings in `/etc/sysconfig/rawdevices` to have rawdevices pointing to these spare partitions

```
$ sudo vi /etc/sysconfig/rawdevices
$ sudo /sbin/service rawdevices restart
```

5. Install CRS as described here

6. If needed upgrade the CRS installation with Patch Set as described here

7. As root stop CRS on both nodes and from one of them (still as root) restore OCR from a backup

```
$ /etc/init.d/init.crs disable
$ /etc/init.d/init.crs stop
$ ps -ef | grep crs
```

If any of the processes exist - kill them.

=$ /ORA/dbs01/oracle/product/10.1.0/crs/bin/ocrconfig -restore ~oracle/ocr_backups/day.ocr

8. If there is no backup of the ASM spfile copy it from other RAC with `'CREATE PFILE FROM SPFILE'` and `'scp'` commands, modify it if necessary and restore it on the '/dev/raw/raw3' device with the `"CREATE SPFILE='/dev/raw/raw3' FROM PFILE='initfile_location'"` command.

9. As root enable CRS on both nodes and restart them

```
$ /etc/init.d/init.crs enable
$ shutdown -r now
```

Everything should startup automatically.

Replacing broken storage and ASM mirror rebuilding

1. Configure spare storage in the same way as the broken storage was configured (logical drives and host LUNs)

2. Go to the computer center, detach the broken storage from the FC switch and attach the new storage to the same port (on both switches). Connect the broken storage to ports where the new storage was connected before. Update the web page.

3. Make new storage visible on OS level and create appropriate partitions if necessary. To add scsi devices on OS level one can use the following command (as root on both nodes):

```
$ echo "scsi add-single-device a b c d" > /proc/scsi/scsi
```

where:
a == hostadapter id (first one being 0)
b == SCSI channel on hostadapter (first one being 0)
c == ID
d == LUN (first one being 0)

In our architecture:
a = 1 (if the new storage is visible through the first port on HBA) or = 2 (if the new storage is visible through the second port on HBA)
b == 0 (always)
c == 0 (always)
d == 0->7 (if there is 8 LUN mappings in the storage)

In a similar way one can remove SCSI disk on-line:

```
$ echo "scsi remove-single-device a b c d" > /proc/scsi/scsi
```

If you don't want to play with commands above you can restart nodes and the result should be the same.

4. As root on one of the cluster nodes clear disks with 'dd' command and create appropriate ASMLIB mappings:

```
dd if=/dev/zero of=/dev/sdj4 bs=1024 count=100000
/etc/init.d/oracleasm createdisk ITSTOR18_1 /dev/sdj4

i=2
for n in k l m n o p q
do
  dd if=/dev/zero of=/dev/sd${n} bs=1024 count=100000
  /etc/init.d/oracleasm createdisk ITSTOR18_${i} /dev/sd${n}
  let i=$i+1
done;

/etc/init.d/oracleasm listdisks
```

On the other nodes discover newly created ASMLIB mappings:

```
$ /etc/init.d/oracleasm scandisks
```

5. Check whether new disks are visible on ASM level. To do this connect to an ASM instance and execute:

```
SQL> SELECT DG.NAME, D.PATH, D.FAILGROUP, D.MOUNT_STATUS, D.HEADER_STATUS, D.MODE_STATUS,
D.STATE FROM V$ASM_DISKGROUP DG, V$ASM_DISK D WHERE DG.GROUP_NUMBER(+) = D.GROUP_NUMBER
ORDER BY 1,3;
```

6. Add disk to diskgroups

```
SQL> ALTER DISKGROUP DATA_DG1 ADD FAILGROUP FG2 DISK 'ORCL:ITSTOR18_3', 'ORCL:ITSTOR18_4',
'ORCL:ITSTOR18_5', 'ORCL:ITSTOR18_6', 'ORCL:ITSTOR18_7', 'ORCL:ITSTOR18_8'; SQL> ALTER
DISKGROUP RECOVERY_DG1 ADD FAILGROUP FG2 DISK 'ORCL:ITSTOR18_2';
```

Note: For some reason (most probably because of bugs) the commands above didn't work for me when ASM instance were started in cluster mode. Also procedure is more reliable where disks are being added one by one.

This topic: PSSGroup > Recovery
Topic revision: r3 - 2005-12-15 - unknown