Die Datenbank-Version 11g R2 hat im Bereich „Real Application Clusters" viel Neues gebracht. Was hat sich hingegen bei der Verwaltung der Clusterware getan? Was macht man, wenn die Diskgruppe GRID nicht mehr zugreifbar ist? Der Fokus dieses Artikels liegt darauf, wie man eine GRID-Installation vornimmt, ein „vollständiges" Backup erstellt und im Falle eines Fehlers die Diskgruppe GRID wiederherstellt.

Diskgruppe GRID weg, Cluster down – was nun?

Stefan Panek, Trivadis GmbH

Die Clusterware beziehungsweise Grid-Infrastruktur ist zentraler Bestandteil der Oracle-RAC-Installation. Sie arbeitet als Bindeglied zwischen Betriebssystem auf der einen und Oracle-Datenbank beziehungsweise ASM-Software auf der anderen Seite. Die Hauptfunktionalitäten sind dabei unter anderem:

- Split Brain Handling
- Ressourcenverwaltung
- Monitoring der Cluster-Ressourcen

Die wichtigsten Komponenten der Grid-Infrastruktur sind (siehe Abbildung 1):

- Voting-Files
- Oracle Cluster Registry (OCR)
- Oracle Local Registry (OLR)
  - Die Oracle Local Registry ist mehr oder minder identisch zur OCR und beinhaltet dabei die lokalen Ressourcen des jeweiligen Clusterknotens
- Grid Plug and Play Profile
  - Grid Plug and Play vereinfacht die Installation, die Konfiguration und das Management des einzelnen Knotens innerhalb eines Clusters

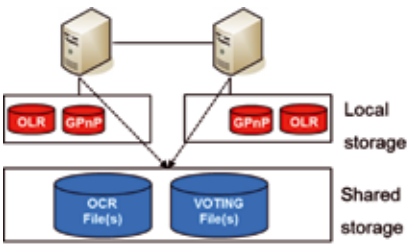

Abbildung 1: Die wichtigsten Komponenten der Grid-Infrastruktur

Die Clusterware-Installation

Die Installation eines Oracle-RAC-Systems besteht aus mehreren, teilweise komplexen Schritten. Im Vorfeld einer Installation ist eine sorgfältige Planung erforderlich. Zu den notwendigen Installationsvorbereitungen zählen die Bereitstellung von IP-Adressen, das Aufspielen von Betriebssystem-Updates und -Packages, die Bereitstellung von genügend Plattenplatz und vieles mehr. Sind diese vorbereitenden Arbeiten abgeschlossen, beginnt die Installation der Grid-Infrastruktur. Der aktuelle Installationsablauf ist im Vergleich zu den Versionen 10*g* R2 oder 11*g* R1 komplett überarbeitet. Wichtige Details werden nun vor der eigentlichen Installation geprüft. Es werden Skripte für das Beheben von Konfigurationsfehlern erstellt, die auch während des Installationsablaufs ausgeführt werden können. Im Folgenden sind einige Teilschritte der gesamten Installation dargestellt.

Nachdem der Installer gestartet wurde, beginnt das GRID Infrastruktur Setup. Nach Prüfung der Knoten-Connectivity sowie einigen Clusterprüfungen werden die Storage-Optionen für das Sichern der OCR-Disks und Voting-Files abgefragt. Dabei gibt es in der Version 11*g* R2 zwei Möglichkeiten:

- Oracle ASM
- Shared File System

In diesem Fall wird ASM als Storage-Option verwendet. Hier werden anschließend auch die OCR Disks & Voting-Files abgelegt. Hinweis: Wird ASM als Storage Option verwendet, so müssen sämtliche Voting-Files dort auch gespeichert sein. Ein Verteilen der Files auf ASM und Shared File System ist zur Installationszeit nicht möglich. Im anschließenden Installationsschritt wird die „Storage-Option" für die Diskgruppe GRID festgelegt. Der Installer bietet dazu drei Varianten an:

- External redundancy
- Normal redundancy
- High redundancy

Hier unterscheidet sich zum Beispiel eine Diskgruppe DATA von einer Diskgruppe GRID. Bei der Diskgruppe DATA bedeu-

Name der Diskgruppe	Redundanz	Failure-Gruppe	Bemerkung
DATA	external	1	Hardware RAID sollte vorhanden sein
DATA	normal	2	
DATA	high	3	
GRID	external	1	Hardware RAID sollte vorhanden sein
GRID	normal	3	
GRID	high	5	

Tabelle 1: Die verschiedenen Optionen der Redundanz

```
clscfg: -install mode specified
Successfully accumulated necessary OCR keys.
Creating OCR keys for user ‚root‘, privgrp ‚root‘..
Operation successful.
CRS-4256: Updating the profile
Successful addition of voting disk 8d44c77clebc4f05bf8de0efd5ac28d3.
Successfully replaced voting disk group with +GRID.
CRS-4256: Updating the profile
CRS-4266: Voting file(s) successfully replaced
##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ----------
 1. ONLINE   8d44c77clebc4f05bf8de0efd5ac28d3  (/dev/sdf) [GRID]
Located 1 voting disk(s).
```

Abbildung 2: Ausgabe des „root.sh“-Skripts

tet „normal“ zwei Fail-Gruppen, bei einer Diskgruppe GRID sind bei „normal“ drei Fail-Gruppen vorhanden. Tabelle 1 zeigt die verschiedenen Optionen der Redundanz. „High“ gibt die konfigurierbare Datenredundanz in der Diskgruppe an. Die empfohlene Redundanz seitens Oracle für die Diskgruppe GRID ist „normal“. Schon zum Zeitpunkt der Storage-Planung ist es sinnvoll, einige „Reserve Disks“ anzulegen. Diese können im Fehlerfall sofort genutzt werden. Somit lässt sich im Ernstfall eine längere Downtime vermeiden. Nachfolgend eine Beispielausgabe der Diskgruppe GRID, konfiguriert mit „normal redundancy“:

```
select GROUP_
NUMBER,FAILGROUP,PATH from
v$asm_disk where group_number
= 3;
GROUP_NUMBER FAILGROUP PATH
------------------------------
3            GRID_0001 /dev/
mapper/FG_02
3            GRID_0002 /dev/
mapper/FG_03
3            GRID_0000 /dev/
mapper/FG_01
```

Man kann hier gut erkennen, dass Oracle die Voting-Files in drei separate Failure-Diskgruppen ablegt. Bei der gewählten „normal redundancy“ sind es somit drei Disks, die hierfür verwendet werden. Der vorletzte Schritt der „Grid Infrastructure“-Software-Installation umfasst das Ausführen des „root.sh“-Skripts. Diese legt die Diskgruppe mit den OCR-Files sowie den Voting-Files an.

Wie Abbildung 2 zeigt, werden sowohl die Diskgruppe als auch die Lokation des Voting-Files angelegt. In diesem Fall wurde „external redundancy“ gewählt. Diese Option ist nur für Testcluster sinnvoll. Nach Abschluss der Installation ist der RAC-Cluster lauffähig und die wesentlichen Ressourcen sind bereits konfiguriert. Damit dies auch so bleibt, muss der Administrator zusätzliche Vorsichtsmaßnahmen treffen. Das notwendige Vorgehen wird im Folgenden näher erläutert.

Planung der Grid-Infrastruktur

So vorteilhaft es ist, alles im ASM abzulegen, so bedarf es einer guten Vorar-

```
OCR Backup
> ocrconfig -showbackup
node-01     2010/11/10 05:57:33     /u01/app/11.2.0.2/grid/cdata/
erac1/backup00.ocr
node-01     2010/11/10 01:57:32     /u01/app/11.2.0.2/grid/cdata/
erac1/backup01.ocr
node-01     2010/11/09 21:57:31     /u01/app/11.2.0.2/grid/cdata/
erac1/backup02.ocr
node-01     2010/11/09 05:57:29     /u01/app/11.2.0.2/grid/cdata/
erac1/day.ocr
node-01     2010/11/01 01:56:56     /u01/app/11.2.0.2/grid/cdata/
erac1/week.ocr
```

Listing 1

```
OLR Check
> ocrcheck -local
Status of Oracle Local Registry is as follows :
         Version                  :          3
         Total space (kbytes)     :     262120
         Used space (kbytes)      :       2712
         Available space (kbytes) :     259408
         ID                       :  905928015
         Device/File Name         : /u01/app/11.2.0.2/grid/
cdata/node-01.olr
                                    Device/File integrity
check succeeded
         Local registry integrity check succeeded
         Logical corruption check succeeded
OLR manual Backup
> ocrconfig -local -manualbackup
```

Listing 2

```
> crsctl query css votedisk

    STATE    File Universal Id                  File Name
Disk group
--  -----    -----------------                  ---------
 1. ONLINE   4191b3802ce14fc2bf41f03d2e3a9df8   (/dev/mapper/
FG1_01)   [GRID]
 2. ONLINE   684686a0c6624f2dbf018c56dd473d7a   (/dev/mapper/
FG1_02)   [GRID]
 3. ONLINE   0b789ea3e8ad4fe4bf346b30abcfa8fb   (/dev/mapper/
FG1_03)   [GRID]
```

Listing 3

beit, um den Cluster im Fehlerfall wieder zum Laufen zu bekommen. Daher gilt es, nach der Installation ein tragfähiges Backup- und Restore-Konzept zu erstellen. Dieses Konzept muss vor Inbetriebnahme ausgiebig getestet werden. Dazu gehört unter anderem auch die Vorgehensweise bei einem Verlust der Diskgruppe GRID:

- Was muss in diesem Fall nun alles wiederhergestellt werden?
- Welche Backups sind dazu vorab durchzuführen?

Nachfolgende Backups sind regelmäßig erforderlich:

- Sichern der Oracle-Software (etwa der Mountpoints)
- OCR-Disks und OLR-Backups auf ein separates Medium kopieren beziehungsweise mit Filesystembackup sichern
- Backup der ASM-Metadaten
- SPFILE der ASM-Instanz

Es empfiehlt sich, über diese Arbeiten eine Dokumentation zu erstellen. Diese sollte als Step-by-Step-Anweisung ausgelegt sein, damit im Fehlerfall das Vorgehen für den Administrator transparent ist. Im Folgenden die Details anhand unseres Beispiels. Ein regelmäßiges Sichern des Oracle Mountpoints ist ein Standard-Job, der hier nicht näher beschrieben wird.

Die Oracle-Software erstellt automatisch im 4-Stunden-Intervall OCR-Backups. Diese können dann mit einem Filesystembackup gesichert werden. Ein Kommando überprüft, ob die Backups regelmäßig erfolgen (siehe Listing 1).

Die Oracle-Software erzeugt automatisch keine OLR-Backups. Ein manuelles Backup ist daher angeraten (siehe Listing 2).

Listing 3 zeigt die Prüfung der Voting-Files. Ein manuelles Backup der Voting-Files ist ab Version 11.2.0.2 nicht mehr notwendig. Diese werden dann automatisch mit dem regelmäßigen OCR-Backup gesichert.

Bei den Metadaten einer ASM-Diskgruppe handelt es sich im Wesentlichen um folgende Informationen:

- Disks, die zu einer Diskgruppe gehören
- Der verfügbare Platz innerhalb der Diskgruppe
- Die File-Namen und der Ort der Files, die zu der Diskgruppe gehören

Das Utility „ASMCMD“ sichert die gesamte Informationen in einer Textdatei. Diese Sicherung kann dann gegebenenfalls für einen Restore benutzt werden:

```
> asmcmd md_backup -b <Pfad>/
Filename
```

Da im Fehlerfall beziehungsweise beim Verlust der Diskgruppe gleichzeitig das SPFILE der ASM-Instanz nicht mehr vorhanden ist, sollte man hier regelmäßig eine Kopie außerhalb von ASM vorhalten. Das Kommando ist wie bei jeder Oracle-Datenbank:

```
SQL> create pfile='<Pfad>/initASM.ora' from spfile
```

Abschließend gilt es, diese Arbeiten zu automatisieren, damit immer die aktuellen Backups verfügbar sind. Dazu bietet sich zum Beispiel unter Unix ein Cronjob an.

Restore der Diskgruppe GRID

Das beste Backup hilft nichts, wenn es vorher nicht getestet wurde. Daher ist vor der Inbetriebnahme des RAC-Clusters unbedingt zu prüfen, ob der Cluster beziehungsweise in diesem speziellen Fall die Diskgruppe GRID wiederhergestellt werden kann. Ein Verlust würde zwangsläufig den kompletten Cluster-Stillstand bedeuten.

Als Ausgangssituation wird angenommen, dass die Diskgruppe GRID korrupt oder vollständig verlorengegangen ist. Nachfolgend ist deren Restore beschrieben.

```
> crsctl start crs -excl
CRS-4123: Oracle High Availability Services has been started.
CRS-2672: Attempting to start ‚ora.mdnsd' on ‚node-01'
CRS-2676: Start of ‚ora.mdnsd' on ‚node-01' succeeded
CRS-2672: Attempting to start ‚ora.gpnpd' on ‚node-01'
CRS-2676: Start of ‚ora.gpnpd' on ‚node-01' succeeded
CRS-2672: Attempting to start ‚ora.cssdmonitor' on ‚node-01'
CRS-2672: Attempting to start ‚ora.gipcd' on ‚node-01'
CRS-2676: Start of ‚ora.cssdmonitor' on ‚node-01' succeeded
CRS-2676: Start of ‚ora.gipcd' on ‚node-01' succeeded
CRS-2672: Attempting to start ‚ora.cssd' on ‚node-01'
CRS-2672: Attempting to start ‚ora.diskmon' on ‚node-01'
CRS-2676: Start of ‚ora.diskmon' on ‚node-01' succeeded
CRS-2676: Start of ‚ora.cssd' on ‚node-01' succeeded
CRS-2672: Attempting to start ‚ora.ctssd' on ‚node-01'
CRS-2672: Attempting to start ‚ora.drivers.acfs' on ‚node-01'
CRS-2672: Attempting to start ‚ora.cluster_interconnect.haip' on
‚node-01'
CRS-2676: Start of ‚ora.ctssd' on ‚node-01' succeeded
CRS-2676: Start of ‚ora.drivers.acfs' on ‚node-01' succeeded
CRS-2676: Start of ‚ora.cluster_interconnect.haip' on ‚node-01' suc-
ceeded
CRS-2672: Attempting to start ‚ora.asm' on ‚node-01'
CRS-2674: Start of ‚ora.asm' on ‚node-01' failed
CRS-2673: Attempting to stop ‚ora.cluster_interconnect.haip' on
‚node-01'
CRS-2677: Stop of ‚ora.cluster_interconnect.haip' on ‚node-01' suc-
ceeded
```

Listing 4

Zunächst ist die Clusterware auf allen Knoten zu „disablen“. Dabei unbedingt auf der Betriebssystemebene prüfen und eventuell einzelne Prozesse mit „kill“ beenden. Dann die Clusterware am Masterknoten im „exclusive mode“ starten.

Die Clusterware versucht nun, die ASM-Instanz zu starten, was aber aufgrund des fehlenden SPFILEs nicht möglich ist. Daher verbindet man sich in einer zweiten Session mit der ASM-Instanz und stoppt diese mit „shutdown abort“ (siehe Listing 4).

Listing 5 zeigt die SQL-Plus-Session zum manuellen Start der ASM-Instanz per PFILE.

Vor dem Restore der Diskgruppe GRID aus den Metadaten muss die ASM-Umgebung gesetzt sein. Nur so kann das Programm „asmcmd“ aufgerufen und genutzt werden (siehe Listing 6).

Zur Prüfung wird die folgende Abfrage durchgeführt:

```
SQL> select name from v$asm_
diskgroup;
NAME
-----------------
GRID
```

Nachdem die Diskgruppe GRID wieder erfolgreich aufgebaut ist, erfolgt der Restore des OCR-Backups:

```
> ocrconfig -restore /u01/
app/11.2.0.2/grid/cdata/erac1/
backup00.ocr
```

Ein Restore des OLR-Backups ist nur notwendig, wenn das OLR verloren wurde. Dieses Backup ist in einem lokalen Verzeichnis des entsprechenden Knotens abgelegt.

Nun wird das SPFILE der ASM-Instanz wieder neu erstellt:

```
SQL> create spfile='+GRID' from
pfile='/backup/initASM.ora;
File created.
```

Ein „replace“-Befehl legt nun die Voting-Files im ASM wieder an (siehe Listing 7). Somit ist der Restore für die Diskgruppe GRID abgeschlossen. Auf dem Restaurierungsknoten wird nun die Clusterware mit der „force“-Option gestoppt:

```
> crsctl stop crs -f (force)
```

Abschließend wird die Clusterware auf allen Knoten gestartet und nochmals zur Kontrolle ein Check durchgeführt (siehe Listing 8). Anschließend ist die Diskgruppe GRID wiederhergestellt und der Cluster funktionstüchtig.

```
$ sqlplus / as sysasm
SQL*Plus: Release 11.2.0.2.0 Production on Thu Nov 11 07:37:32 2010
Copyright (c) 1982, 2010, Oracle.  All rights reserved.
Connected.
SQL> shutdown abort
ASM instance shutdown
SQL> exit
Disconnected
$ sqlplus / as sysasm
SQL*Plus: Release 11.2.0.2.0 Production on Thu Nov 11 07:39:04 2010
Copyright (c) 1982, 2010, Oracle.  All rights reserved.
Connected to an idle instance.
SQL> startup nomount pfile='/backup/init+ASM1.ora'
ASM instance started
Total System Global Area  283930624 bytes
Fixed Size                  2225792 bytes
Variable Size             256539008 bytes
ASM Cache                  25165824 bytes
```

Listing 5

```
asmcmd md_restore  /backup/asm/diskgroup_metadata.lst  --full  -G
GRID
Current Diskgroup metadata being restored: GRID
Diskgroup GRID created!
System template XTRANSPORT modified!
System template ONLINELOG modified!
System template DATAGUARDCONFIG modified!
System template AUTOBACKUP modified!
System template TEMPFILE modified!
System template OCRFILE modified!
System template ARCHIVELOG modified!
System template DUMPSET modified!
System template CONTROLFILE modified!
System template BACKUPSET modified!
System template ASMPARAMETERFILE modified!
System template FLASHBACK modified!
System template PARAMETERFILE modified!
System template FLASHFILE modified!
System template DATAFILE modified!
System template CHANGETRACKING modified!
Directory +GRID/erac1 re-created!
Directory +GRID/erac1/OCRFILE re-created!
Directory +GRID/erac1/ASMPARAMETERFILE re-created!
```

Listing 6

Fazit

Die Clusterware ist das Herzstück einer jeden Real-Applications-Umfeld-Clusters-Installation. Seit der Oracle Version 11g R2 wurde die Clusterware stark überarbeitet und beinhaltet nun deutlich mehr Funktionalitäten. Da seitdem die Clusterkonfiguration oftmals in ASM abgelegt wird, sind im Vorfeld hinreichende Sicherheitsmaßnahmen zu treffen, für den Fall, dass die

Diskgruppe GRID verloren geht. Mit einem entsprechenden „Kochrezept“ ist ein solcher Verlust oder auch eine entstandene Korruption recht schnell und sicher wieder zu beheben. Die Ausfallzeit des Clusters wird so stark reduziert.

Im Rahmen dieses Artikels ist auch der Teilaspekt „Backup & Restore“ der Grid-Infrastruktur beschrieben. Genau diesen Punkt sollte man im Vorfeld einer Cluster-Inbetriebnahme umfassend testen. Für einen stabilen Start in den Datenbankbetrieb sind aber noch weitere Bereiche zu prüfen, wie die Datenbankserver-Konfiguration, der Interconnect, das Netzwerk und weitere Komponenten.

Für einen gesicherten Start in den Betrieb ist eine ausführliche Cluster-Prüfung durchzuführen, bei der alle Bereiche des Clusters ausgiebig getestet und Fehler beziehungsweise eventuell fehlerhafte Konfigurationen angepasst werden. So können Systeme auch in Zukunft hochverfügbar bleiben.

Stefan Panek
Trivadis GmbH
stefan.panek@trivadis.com

```
> crsctl replace votedisk +GRID

Successful addition of voting disk 3ae1d7ce99014f79bf06ae8e039f8d34

Successful addition of voting disk 0a21e3b2d11b4fe2bf6a7b175777808a

Successful addition of voting disk 9d762eab80254fdabf65ffe67801cc98

Successfully replaced voting disk group with +GRID.
CRS-4266: Voting file(s) successfully replaced


Prüfen der Voting Files

> crsctl query css votedisk

##  STATE    File Universal Id                File Name Disk group
--  -----    -----------------                --------- ---------
 1. ONLINE   3ae1d7ce99014f79bf06ae8e039f8d34 (/dev/mapper/
FG1_01) [GRID]
 2. ONLINE   0a21e3b2d11b4fe2bf6a7b175777808a (/dev/mapper/
FG1_02) [GRID]
 3. ONLINE   9d762eab80254fdabf65ffe67801cc98 (/dev/mapper/
FG1_03) [GRID]
Located 3 voting disk(s).
```

Listing 7

```
crsctl start crs
crsctl check cluster -all
**************************************************************
node-01:
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
**************************************************************
node-02:
CRS-4537: Cluster Ready Services is online
CRS-4529: Cluster Synchronization Services is online
CRS-4533: Event Manager is online
**************************************************************
```

Listing 8

Impressum

Herausgeber:
DOAG Deutsche ORACLE-Anwendergruppe e.V.
Tempelhofer Weg 64, 12347 Berlin
Tel.: 0700 11 36 24 38
www.doag.org

Verlag:
DOAG Dienstleistungen GmbH
Fried Saacke, Geschäftsführer
info@doag-dienstleistungen.de

Chefredakteur (ViSdP):
Wolfgang Taschner
redaktion@doag.org

Chefin von Dienst (CvD):
Carmen Al-Youssef
office@doag.org

Titel, Gestaltung und Satz:
Claudia Wagner, Katja Borgis
DOAG Dienstleistungen GmbH

Titelfoto: Fotolia

Anzeigen:
CrossMarketeam Ralf Rutkat, Doris Budwill
www.crossmarketeam.de
Mediadaten und Preise finden Sie unter:
www.doag.org/publikationen/

Druck:
adame Advertising and Media GmbH Berlin, www.adame.de