Oracle Clusterware 11*g* R2

Martin Bach, Martin Bach Consulting Ltd.

Dieser Artikel zeigt einige neue Features von Oracle Clusterware 11*g* R2. Insbesondere die Neuerungen und Unterschiede zu der verbreiteten Vorversion 10*g* R2 erfordern von den Administratoren einige Einarbeitungszeit. Die bereits bekannte und gut dokumentierte Startsequenz der Hintergrunddienste hat sich in wesentlichen Punkten verändert und verdient gesonderte Betrachtung. Manche Administratoren wissen auch nicht, dass Clusterware neben RAC ebenso zum Aufbau von Aktiv/Passiv-Clustern geeignet ist. Kenntnisse der Versionen 10*g* und 11*g* R1 sind für das Verständnis des Artikels empfehlenswert.

Oracle 11*g* R2 bot dem interessierten Administrator einige Überraschungen. Angefangen damit, dass die ZIP-Dateien für die Datenbank-Installation bereits fast 2 GB umfassten, bis hin zur Einführung von Oracle Restart und Grid Infrastructure musste sich der Nutzer an einige Änderungen gewöhnen. Allen, die so wie der Autor lieber erst einmal ein neues Release ausprobieren möchten, ohne gleich das „New Features Guide" bis ins Detail zu studieren, war recht schnell klar, dass diesmal ohne Studium der Handbücher nicht viel zu gewinnen ist.

Oracle Clusterware im Kontext

Oracle Clusterware, oder Grid Infrastructure, wie es im folgenden genannt wird, ist die Basis für jeden Real Application Cluster. RAC 11*g* R2 besteht aus den beiden Software-Komponenten Grid Infrastructure und Datenbank-Software. Leser, die bereits Erfahrung mit RAC in den Releases 10*g* R2 und 11*g* R1 gemacht haben, werden das separate ORACLE_HOME für ASM vermissen. Vor 11*g* R2 riet Oracle dazu, ASM ein eigenes ORACLE_HOME zu spendieren. Dies diente vor allem zwei Zwecken:

1. Die System-Administratoren konnten einen Teil ihrer Befugnisse zurückbekommen, indem das Management von ASM von der Datenbank entkoppelt war. Oracle bediente sich bis einschließlich 11*g* R2 dem Konzept unterschiedlicher Betriebssystem-Konten. So fand sich auf manchen Systemen ein Benutzer „asm", unter dessen Konto die ASM-Software installiert wurde. Die Datenbank selbst war dann wieder unter dem Benutzer „oracle" installiert. Die Zuweisung von Betriebssystem-Gruppen zu den Rollen OSASM, OSOPER und OSDBA garantierten die Aufgabenteilung.
2. ASM konnte unabhängig von der Datenbank zu einer neuen Version migriert werden. Sofern ASM dasselbe ORACLE_HOME wie die Datenbank verwendete, mussten beide gleichzeitig auf die neue Version migriert werden. Eine Entkopplung der Software in unterschiedliche ORACLE_HOMEs ermöglichte es zum Beispiel, Clusterware und ASM von 10*g* R2 zu 11*g* R1 zu migrieren und gleichzeitig die Datenbank zu belassen.

In Oracle 11*g* R2 ist dieses Konzept überholt: ASM ist nicht mehr Bestandteil der Datenbank-Installation, sondern Teil der Infrastruktur-Schicht. Dies bedeutet, dass ASM und die Clusterschicht eine Einheit bilden und nicht mehr getrennt aktualisiert werden können. Eine Sorge-Rollenteilung ist immer noch möglich, wie später beschrieben wird. Abbildung 1, entnommen aus Shaw/Bach „Pro Oracle Database 11*g* RAC on Linux", verdeutlicht das neue Konzept.

Oracle Cluster Layer

Wie eingangs beschrieben, ist Clusterware die Basis für RAC. Neben RAC bieten sich aber noch weitere mögli-

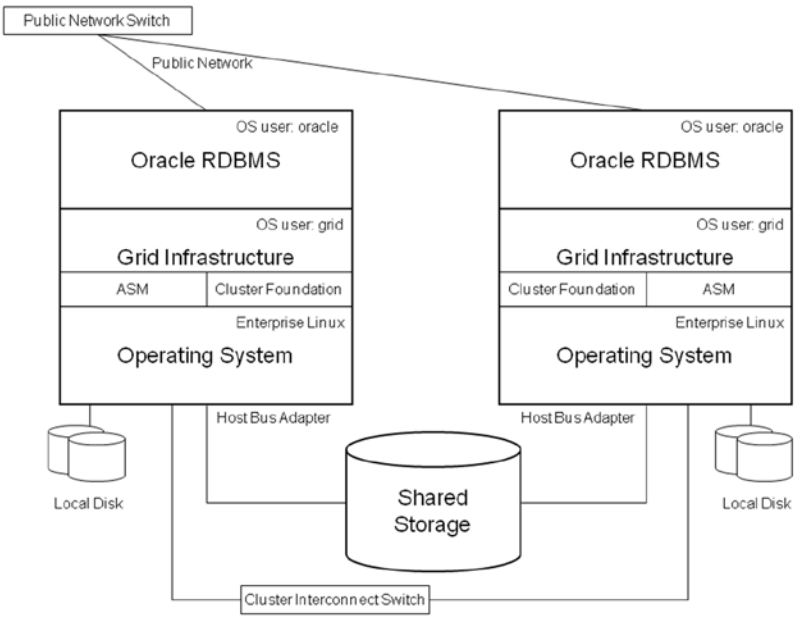

Abbildung 1: Das neue Konzept der Oracle Clusterware

che Aufgabenfelder für Clusterware. Abstrahiert man von RAC, bietet sich ein spannendes Bild, da Clusterware ein vollständiges HA-Framework darstellt. Es ist ohne Weiteres möglich, eigene Ressourcen in Clusterware zu definieren und diese von Clusterware verwalten zu lassen. Clusterware bietet dabei noch ein POSIX Cluster-File-System (ACFS) für jegliche Binärdateien und einen Cluster-Logical-Volume-Manager in Form von ASM. Das Profil der Ressource bestimmt, wie Clusterware im Fehlerfall vorgehen soll. Der Autor hat bereits erfolgreich mehrere Aktiv/Passiv-Cluster allein mit Clusterware implementiert, die eine „single instance" Oracle-Datenbank oder Netzwerkdienste wie Apache Tomcat vor dem Ausfall eines Cluster-Knotens schützen.

Oracle wildert damit in den Gefilden, die den einstigen Cluster-Pionieren vorbehalten waren. Dabei ist Oracle-Clusterware weiterhin kompatibel zu anderen, nicht-Oracle Cluster-Managern wie HACMP von IBM oder Veritas. Mit der von Clusterware zur Verfügung gestellten Fülle an Features stellt sich allerdings die Frage nach dem Sinn einer Installation von externer Software. Nach Erfahrung des Autors entstehen nicht nur zusätzliche Lizenzkosten, sondern auch eine Situation, in der es für den Support von Oracle und dem externen Software-Anbieter ein Einfaches ist, dem jeweils anderen die Verantwortung für ein Problem in die Schuhe zu schieben. Für bestimmte Plattformen, die eine Bündelung beispielsweise von Netzwerkschnittstellen nicht mit Bordmitteln ermöglichen, ist eine Erweiterung des Software Stack allerdings oft unausweichlich.

Absicherung der Oracle-Datenbank in einem Aktiv/Passiv-Cluster

Die Absicherung der Datenbank ist mit einem Cluster recht einfach. Je nach Einsatzzweck kann ein solcher Cluster sogar die Hochverfügbarkeitsanforderungen erfüllen und somit eine RAC-Lizenz ersparen. Die Installation der Software ist unkompliziert – wie in einer typischen RAC-Installation wird Grid Infrastructure auf beiden Cluster-Knoten installiert. Es sollte zudem eine ASM-Disk-Group für die Speicherung der Datenbank zum Einsatz kommen. Wichtig ist nur, dass man im nächsten Schritt eine single-instance RDBMS-Installation vornimmt. Entweder kann diese auf jedem Cluster-Knoten erfolgen oder, im Fall der Verwendung von ACFS, in einem gemeinsamen ORACLE_HOME.

Nach der Erstellung der Datenbank – der Einfachheit halber wird angenommen, sie sei nicht mit DBCA erstellt worden – muss diese als Ressource in Clusterware registriert werden. Zu dieser Registrierung ist ein sogenanntes „Action-Script" notwendig, das die Callback-Funktionen „start()", „stop()", „clean()" und „check()" der Cluster ware ausführt. Dieses Skript könnte SQL-Plus aufrufen, um die Datenbank zu starten und zu stoppen, und in der „check()"-Funktion das Vorhandensein des PMON-Prozesses überprüfen. Dieses Action-Skript muss auf beiden Cluster-Knoten in einem für den Grid-Infrastructure-Nutzer erreichbaren Verzeichnis liegen. Dann lässt sich die von Clusterware zu überwachende Ressource definieren, am einfachsten mit einem Property-File:

```
PLACEMENT=restricted
HOSTING_MEMBERS=node1 node2
CHECK_INTERVAL=30
CARDINALITY=1
ACTIVE_PLACEMENT=0
AUTO_START=restore
DEGREE=1
DESCRIPTION="Oracle Database
Resource"
RESTART_ATTEMPTS=1
ACTION_SCRIPT=/u01/app/crs/ha-
daemon/hacluster.sh
```

In diesem Beispiel ist das Action-Skript als „hacluster.sh" hinterlegt. Die Registrierung erfolgt dann mit dem Aufruf von „crsctl add resource". Sobald die Ressource mit dem Root-Benutzer registriert wurde, sollten ACLs gesetzt werden, die dem Grid- und RDBMS-Benutzer gestatten, die Ressource mit „crsctl (start | stop)" zu starten oder zu stoppen. Die ACLs werden mit „crsctl setperm" und „crsctl getperm" gesetzt oder abgefragt.

Clusterware-Prozesse in zwei Klassen eingeordnet

Bis einschließlich Oracle 11*g* R1 bestand die Cluster-Software hauptsächlich aus den drei durch „init" gestarteten Hintergrund-Prozessen Cluster Ready Services (CRS), Cluster Synchronisation Services (CSS) und Event Manager (EVM). Hinzu kamen der Oracle Notification Service (ONS) und der Process Monitor Daemon (OPROCD). Die Prozesse des RACG-Stapels kümmerten sich in diesem Fall um das Resource-Management.

Die meisten davon existieren auch weiterhin in Grid Infrastructure, aber die Art und Weise, wie sie verwendet werden, hat sich geändert. Oracle hat die Clusterware-Prozesse in zwei Klassen eingeordnet: den sogenannten „High Availability Stack (HA)" und den „Cluster Ready Stack (CRS)". Der HA-Stack bedient die Grundbedürfnisse eines Clusters, also Dienste, die zwingend zur Kommunikation der Cluster-Knoten miteinander erforderlich sind. Erst mit dem erfolgreichen Start von CRSD kann ein Knoten Teil des Clusters werden. Der CRS übernimmt von dort an und startet die Cluster-Ressourcen wie ASM, die Datenbanken sowie abhängige Dienste, Netzwerkkomponenten etc.

Anstatt CRSD, CSSD und EVMD direkt per „inittab" zu starten, übernimmt der neue Oracle High Availability Service Daemon (OHASD) das Kommando. Dieser Prozess ist der einzige Eintrag in „/etc/inittab" und verantwortlich für den Start der weiteren Komponenten. OHASD hat nach seinem erfolgreichen Start die Aufgabe, die sogenannten „Agent-Prozesse" zu starten. Diese sind neu in 11*g* R2 und erfüllen vielfältige Aufgaben; vereinfacht gesagt übernehmen sie die Funktionen, die vormals von den RACG-Prozessen erfüllt wurden. Etwas Verwirrung stiften zwei völlig verschiedene Agent-Prozesse: einmal die gerade angesprochenen OHASD-Agent-Prozesse, sowie die später beschriebenen CRSD-Agent-Prozesse. Die von OHASD gestarteten Agent-Prozesse mit ihren Aufgaben sind in der Tabelle 1 aufgeführt.

OHASD-Agent-Prozess	Aufgabe(n)
ORAROOTAGENT	Startet Ressourcen, die root-Rechte auf dem Betriebssystem erfordern. Dazu zählen: • ASM Cluster File System und Treiber • Diskmon • Cluster Time Sync Daemon • CRSD (wichtig!)
ORAAGENT	Startet Ressourcen, die keine root-Rechte benötigen. Unter anderem diese: • Multicast DNS, verwendet zur Namensauflösung • ASM • Event Manager • Grid Plug and Play • Grid Inter Process Communication
CSSDMONITOR	Überwacht den CSSD-Prozess
CSSDAGENT	Erzeugt den CSSD-Prozess und überwacht gemeinsam mit CSSDMONITOR dessen Status

Tabelle 1: Die Aufgaben der von OHASD gestarteten Agent-Prozesse

Nach dem Start der Agenten tritt der Cluster in die nächste Startphase ein. Ein funktionierender CRSD erlaubt es dem Knoten, dem Cluster beizutreten. CRSD startet, wie bereits angesprochen, seine eigenen Agent-Prozesse: wieder einen ORAROOTAGENT und einen ORAAGENT. Diese sind nun, vereinfacht dargestellt, zuständig für Cluster-Dienste. Die von ORAAGENT gestarteten Dienste hängen davon ab, ob es sich um eine RAC-Installation oder einen Aktiv/Passiv-Cluster handelt. Im Folgenden wird davon ausgegangen, dass es sich um eine RAC-Installation handelt. Der ORAAGENT-Prozess startet in diesem Fall die ASM-Instanz, die ASM Disk Groups, die RDBMS-Instanzen, die weiter unten beschriebenen „Single Client Access Name Listener" und die virtuellen IP-Adressen. Des Weiteren werden die Datenbankdienste („Services") sowie der Oracle Notification Service gestartet.

Rollenteilung in RAC 11*g* R2

In der Einleitung war die Rede von der Bündelung von ASM und Clusterware in einem ORACLE_HOME in der Grid Infrastructure. Eine Installation von ASM unter einem dedizierten Benutzerkonto ist damit nicht möglich – ASM und Clusterware teilen sich ein ORACLE_HOME. Dennoch ist eine sehr fein granulare Rollenteilung möglich. Zuerst einmal sollte Grid Infrastructure mit einem von den Oracle-Binaries verschiedenen Benutzer installiert werden. Oftmals wird der Benutzer „grid" dazu verwendet. Wichtig ist, dass dieser eigene Betriebssystemgruppen hat für die OSASM-, OSDBA- und OSOPER-Rollen – eine globale Zuweisung von „oinstall" oder „dba" ist nicht empfehlenswert, sofern eine Aufgabenteilung erwünscht ist. Die Datenbank-Installation wiederum sollte mit einem weiteren Benutzer erfolgen, der ebenso eigene Gruppen für die OSOPER- und OSDBA-Gruppen besitzt und auf OS-Ebene Mitglied der Gruppe für OSASM ist (ansonsten hätte er keinen Zugriff auf die ASM-Instanz).

Zudem erlaubt Grid Infrastructure das Anlegen von Cluster-Administratoren. Diese können Zugriff auf bestimmte Cluster-Ressourcen in Form von „Access Control Lists" bekommen. Dieses Feature muss explizit aktiviert und konfiguriert werden, es dient vor allem dem neuen Konzept „Quality of Services", das mit 11.2.0.2 hinzugekommen ist.

Single Client Access Name (SCAN)

SCAN-Adressen sind weitere virtuelle IP-Addressen, die Grid Infrastructure zusätzlich zu den Node-VIPs verwendet. Bevor eine Installation von Grid Infrastucture erfolgen kann, muss die SCAN im Domain-Name-System eingetragen sein. Dazu muss der Name, üblicherweise „*clustername*-scan" zu mindestens einer und höchstens drei IP-Adressen auflösbar sein. Nach der Installation von Grid Infrastructure stehen dann ein bis drei neue virtuelle IP-Adressen zur Verfügung, die über alle Knoten des Clusters verteilt werden. Die SCAN repräsentiert den Cluster als solchen, nicht die Datenbank, die auf ihm läuft. Die Verwendung der SCAN abstrahiert auf elegante Weise von der Anzahl der Cluster-Knoten. Außerdem dienen die SCAN-VIPs und die dazugehörigen Listener, von denen es maximal drei gibt, dazu, die lokalen Listener zu entlasten. In RAC 11*g* R1 und früher musste in der ADDRESS_LIST jede Cluster-VIP angegeben werden, wie das folgende Beispiel zeigt:

```
prod=
(DESCRIPTION=
  (ADDRESS_LIST=
    (LOAD_BALANCE = on)
(FAILOVER = ON)
    (ADDRESS=(PROTOCOL=tcp)
(HOST=node1-vip)(PORT = 1521))
    (ADDRESS=(PROTOCOL=tcp)
(HOST=node2-vip)(PORT = 1521))
    (ADDRESS=(PROTOCOL=tcp)
(HOST=node3-vip)(PORT = 1521))
    (CONNECT_DATA =
      (SERVICE_NAME = prod))))
```

Mit der Verwendung der SCAN ist dies sehr vereinfacht möglich:

```
prod=
(DESCRIPTION=
  (ADDRESS_LIST=
    (ADDRESS=(PROTOCOL=tcp)
(HOST=cluster-scan)(PORT =
1521))
    (CONNECT_DATA =
    (SERVICE_NAME = prod))))
```

Der Vorteil liegt auf der Hand: Wird der Cluster um einen weiteren Knoten ergänzt, müssen keinerlei lokale tnsnames.ora-Dateien zur Namensauflösung angepasst werden, um clientseitiges Load Balancing zu gewährleisten.

Voting Disks und Oracle Cluster Registry

Seit der Einführung von Grid Infrastructure ist es möglich, „voting files" und die Oracle Cluster Registry (OCR) in ASM zu hinterlegen. In Versionen

bis 11*g* R2 waren die folgenden Speicherorte für diese kritischen Dateien erlaubt:

- Raw device
- Block device (seit 10*g* R2)
- Ein unterstütztes Cluster File System (ocfs2 und andere)

Der Vorteil der Speicherung von „voting files" und der OCR in ASM ist hauptsächlich die einfachere Handhabung dieser Dateien im Vergleich zu Raw- und Block-Devices. Es ist keine Sonderbehandlung jener Dateien in „udev" nötig, und Benutzer von ASMLib können endlich ASM-Disks und „OCR/voting files" auf die gleiche Art adressieren. Der Verbleib von OCR und „voting files" in Raw-/Block-Devices ist übrigens nur für migrierte Systeme zulässig; Neuinstallationen können nur zwischen CFS und ASM wählen. Nach Ansicht des Autors ist die Verwendung eines CFS etwas problematisch, da die meisten eine zusätzliche Cluster-Schicht (heartbeat, IO-fencing etc.) mit sich bringen und Cluster-Knoten aus dem Cluster auszuschließen vermögen, ohne dass Oracle Clusterware dasselbe getan hätte.

Neues in Automatic Storage Management

ASM hat ebenso einige Neuerungen erfahren, die interessantesten sind nach Ansicht des Autors die folgenden:

- ASM Cluster File System: Ein POSIX-kompatibles Dateisystem, das „copy-on-write"-Snapshots und seit 11.2.0.2 Replikation beherrscht. Mit Ausnahme von Oracle-Datafiles kann ACFS so ziemlich alles an Dateien speichern und auch als Ziel einer Datenbank-Installation als gemeinsames ORACLE_HOME dienen. Clusterweite „external tables" und Directory-Objekte werden endlich Wirklichkeit.
- Access Control Lists: ASM Disk Groups können so konfiguriert sein, dass Nutzer mit SYSDBA nur bestimmte, ihnen gehörende Dateien verändern können. Dies soll versehentliches Löschen von Dateien in gemeinsam genutzten Disk Groups verhindern.
- Intelligent Data Placement: Interessant wohl nur für DAS oder Exadata, erlaubt das IDP die Aufteilung einer ASM Disk in „hot"- und „cold"-Regionen. Mittels eines Templates lassen sich Tablespaces und Control Files in die äußeren Sektoren einer ASM Disk verfrachten. Für Storage Area Networks, die ASM Disks bereitstellen, ist dies wohl keine Option – einzelne LUNs sind dort normalerweise viele Festplatten, über die Daten verteilt werden. „hot"- und „cold"-Regionen ergeben dort keinen Sinn, da diese LUNs keine äußeren und inneren Sektoren im Sinne einzelner Festplatten haben.
- Unterstützung für Sektorgrößen > 512 Byte. Sofern alle ASM Disks dies unterstützen, kann eine ASM Disk Group mit Sektorgrößen von 4 KB anstatt 512 Byte erstellt werden. Viele Platten sind heute schon mit mehr als zwei TB zu bekommen, aber die Adressierung mit 512-Byte-Sektoren erscheint dabei fast schon archaisch.
- Das neue Tool „renamedg" erlaubt es endlich, eine bestehende Disk Group umzubenennen.

Fazit

Der vorliegende Artikel kann natürlich nur die Spitze des Eisbergs sein, Clusterware ist ein enorm umfangreiches Thema. Zudem ist die Entwicklung von Clusterware ständig in Bewegung, und es scheint, als würde Oracle, anstatt nur zu patchen, auch wesentliche neue Funktionen in Point-Releases verteilen. Patchset 1 für Oracle 11.2.0.2 hat 34 neue Features und ein eigenes Kapitel im „New Features Guide" bekommen – es bleibt also spannend.

Kontakt:
Martin Bach
martin@martinbach-consulting.com