Oracle RAC 11.2 mit Standby-System (Data Guard) an entferntem Standort

Susanne Jahr
Herrmann & Lenz Services GmbH
Burscheid

Schlüsselworte:

RAC, 11gR2, Data Guard, Solaris, NetApp

Einleitung

Die pharma mall GmbH, Betreiber einer Web-Anwendung zur Bestellung von Medikamenten für Krankenhäuser und Apotheken, benötigt für ihre Oracle Datenbanken ein ausfallsicheres System. Es existierte bereits ein Real Application Cluster Oracle 10gR2 unter Solaris x64 bei einem externen Hosting-Anbieter. Nunmehr sollen die Datenbank-Server jedoch im eigenen Rechenzentrum in Sankt Augustin-Hangelar betrieben werden. Da in der bereits vorhandenen Datenbank-Serverlandschaft Solaris x64 erfolgreich als Betriebssystem eingesetzt wird, soll dies auch im neuen 2-Knoten-RAC der Version 11.2 beibehalten werden. Zur zusätzlichen Absicherung wird eine physikalische Standby-Datenbank mit Oracle Data Guard eingerichtet, deren Standort das Backup-Rechenzentrum im ca. 40km entfernten Köln ist.

Betriebssystem, Hardware, Shared Storage

Das eingesetzte Betriebssystem Solaris 10 läuft auf zwei SUN Fire X42xx mit je einem AMD-Quad-Core Prozessor und 32 GB RAM. Jeder Server verfügt über zwei 300 GB große gespiegelte SAS Festplatten.

Das Shared-Storage der RAC-Installation befindet sich auf einem NetApp FAS2040 File-Server mit Cluster-Controllern und redundanter Netzwerkanbindung. Die angebundenen Harddisk-Shelves werden von je einem Controller bedient und sind geclustert.

Da NetApp Filer clusterfähige, von Oracle unterstützte Dateisysteme per NFS zur Verfügung stellen können, wurden diese direkt als Basis für den Shared Storage verwendet. Hier wurden sowohl die Grid Infrastructure-relevanten Dateien (Voting Disks und OCR) gespeichert als auch die Datenbankdateien selbst. Auf eine Zwischenschicht mit ASM wurde daher entgegen der ursprünglichen Planung verzichtet. Da insgesamt drei physikalisch unabhängige Shared-Storage-Speicherorte vorhanden sind, kann jede der drei Voting Disks in einem eigenen, unabhängigen Standort gespeichert werden. So wird an dieser Stelle das Dilemma des Single Point of Failure bezüglich der Voting Disk verhindert. Dieses entsteht dann, wenn man drei Voting Disks auf zwei SAN-Türme verteilen muss, wobei jedoch zur Lauffähigkeit des Clusters mindestens zwei der drei konfigurierten Voting Disks jederzeit für alle RAC-Knoten verfügbar sein muss.

Netzwerk

Die Verbindung der beiden Rechenzentren erfolgt über redundant ausgelegte Cisco3750-Switches. Es steht eine doppelte 10Gbit-Leitung zur Verfügung. Sowohl die Switches zwischen den beiden RAC-Knoten für den Cluster Interconnect als auch diejenigen für das öffentliche Netzwerk sind redundant ausgelegt. Die physikalischen Netzwerkkarten für das private Netzwerk sind auf jedem Knoten redundant ausgelegt und mittels Link Aggregation zu jeweils einer virtuellen Schnittstelle verbunden.

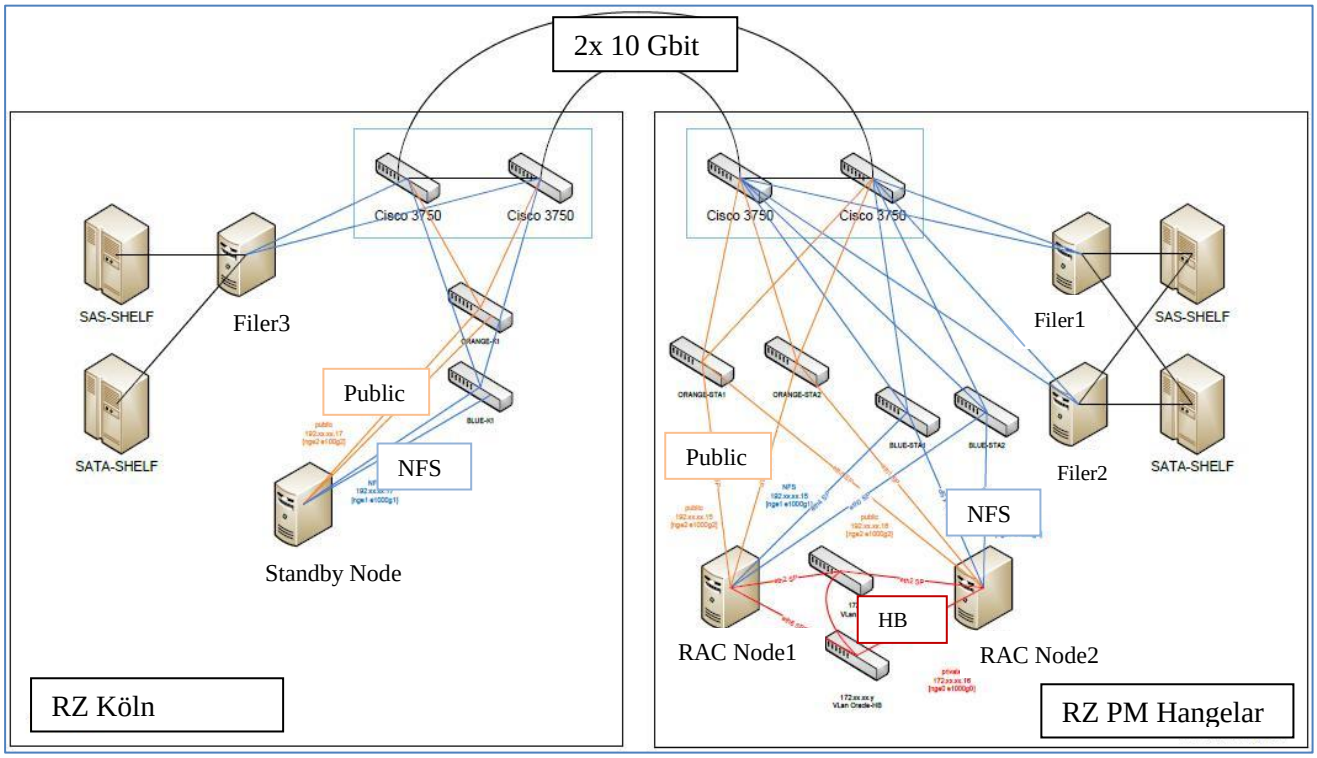

Abb. 1: Schaubild RAC-System pharma mall GmbH

Installation / Konfiguration des RAC

Die Installation der Grid Infrastructure und der Datenbank-Software erfolgte von RAC-Node1 aus per Oracle Universal Installer.

Die für die OCR und Voting Disk benötigten speziellen Mount-Optionen der NFS-Shares wurden gemäß *Oracle Grid Infrastructure Installation Guide for Solaris Operating System* konfiguriert (Beispiel: NFS-Share für die 3 verschiedenen OCR; die jeweiligen Host-Namen wurden mit *filerSTA1* und *filerSTA2* für die beiden NetApp-Filer im primären Standort Sankt Augustin und mit *filerK* für den im Backup-RZ Köln ersetzt):

```
<filerSTA1>:/vol/oraclevdocr - /mnt/oracleasm/oraclevdocr nfs - yes
rw,bg,hard,nointr,rsize=32768,wsize=32768,proto=tcp,vers=3,noac,forcedirectio

<filerSTA2>:/vol/oraclevdocr2 - /mnt/oracleasm/oraclevdocr2 nfs - yes
rw,bg,hard,nointr,rsize=32768,wsize=32768,proto=tcp,vers=3,noac,forcedirectio

<filerK>:/vol/oraclevdocr3 - /mnt/oracleasm/oraclevdocr3 nfs - yes
rw,bg,hard,nointr,rsize=32768,wsize=32768,proto=tcp,vers=3,noac,forcedirectio
```

Im neu gewandeten Oracle Universal Installer können die entsprechenden File-Systeme dann direkt verwendet werden:

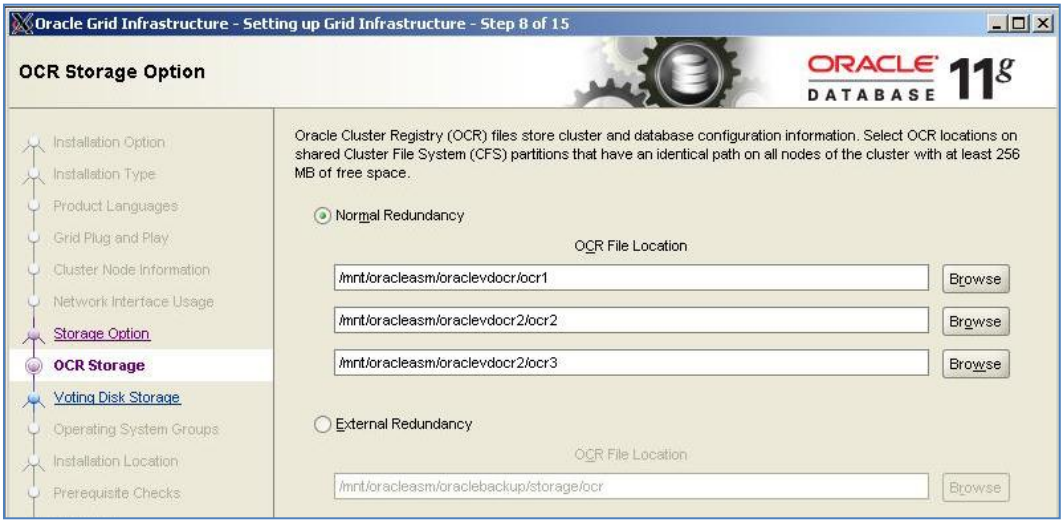

Abb. 2: Auswahl der NFS-Shares für die OCR (Universal Installer)

Die Mount-Optionen für die Nicht-Grid-Infrastructure-Dateien (exemplarisch für das Verzeichnis der Datenbank-Dateien) unterscheiden sich geringfügig:

```
<filerSTA1>:/vol/oracledata - /mnt/oracleasm/oracledata nfs - yes
rw,bg,hard,nointr,rsize=32768,wsize=32768,proto=tcp,noac,forcedirectio,vers=3,suid
```

Der Installer in 11.2 kann außerdem auch die ssh-Konnektivität zwischen den beteiligten Knoten konfigurieren und eventuell fehlende Kernel-Parameter in ein Skript verwandeln, das man dann als root-User ausführen kann, ohne die Installer-Session neu zu starten. Der 11.2-Installer kann also, was die Arbeitserleichterung bei RAC-Installationen angeht, als sehr gelungen bezeichnet werden.
Die Erstellung und Konfiguration der Cluster-Datenbanken (insgesamt laufen hier drei) erfolgte per Skript – wahlweise kann natürlich auch der dbca verwendet werden. Es sind also pro Datenbank zwei Instanzen (auf jedem RAC-Knoten eine) vorhanden.

Änderungen gegenüber Pre-11.2-Systemen – die Grid Infrastructure

Abgesehen vom neuen Installer in 11.2 hat sich auch einiges im gesamten Aufbau des RAC getan. An die Stelle der bekannten Clusterware ist die Grid Infrastructure getreten, die im Gegensatz zu vorherigen Versionen nicht nur die zum Betrieb des Clusters benötigten Binaries sowie OCR und Voting Disk beinhaltet, sondern auch das Automatic Storage Management (ASM)-System sowie den Listener. Die genannten Komponenten teilen sich ein gemeinsames ORACLE_HOME, das nunmehr als Grid Infrastructure Home bezeichnet wird. Die Speicherung der OCR und Voting Disk(s) auf Raw Partitions wird für neue Installationen nicht mehr unterstützt (allerdings schon für Upgrades älterer Systeme). Diese Dateien können entweder (falls verwendet) in ASM, auf einem unterstützten Cluster-File-System oder – wie im vorliegenden Projekt – auf NFS-Shares unterstützter Anbieter wie z.B. NetApp gespeichert werden.

Eine weitere Neuerung in der Version 11.2 ist der virtuelle, clusterweit gültige Name für den Verbindungsaufbau durch die Clients an Datenbanken im Cluster. Für diesen Namen (**S**ingle **C**lient **A**ccess **N**ame, SCAN) wurden nach Empfehlung des *Installation Guides* im DNS drei IP-Adressen

(sog. SCAN-VIPs) hinterlegt, die im Round-Robin-Verfahren abwechselnd bei Ansprache des SCAN aufgelöst werden.

Die bei der Installation der Grid Infrastructure ebenfalls erzeugten drei SCAN-Listener sind zwischen die Clients und die lokalen Listener geschaltet und kennen alle im Cluster registrierten Services. Jeder der SCAN-Listener (LISTENER_SCAN1, 2, 3) hört auf einer der SCAN-VIP-Adressen. Eine Client-Verbindung wird somit unter Verwendung lediglich des SCAN anstelle der bisher im RAC verwendeten Host-Adresslisten auf einen der drei SCAN-Listener treffen und von diesem an denjenigen lokalen Listener weitergeleitet werden, der zur Zeit am wenigsten Last verarbeiten muss.

Abfragen über Konfiguration und Status der SCAN-Listener können, wie gewohnt, mit dem `srvctl`-Kommando durchgeführt werden:

```
oracle@RAC-Node1:~> srvctl status scan_listener
SCAN Listener LISTENER_SCAN1 is enabled
SCAN listener LISTENER_SCAN1 is running on node RAC-Node1
SCAN Listener LISTENER_SCAN2 is enabled
SCAN listener LISTENER_SCAN2 is running on node RAC-Node2
SCAN Listener LISTENER_SCAN3 is enabled
SCAN listener LISTENER_SCAN3 is running on node RAC-Node1

oracle@ RAC-Node2:~> srvctl config scan
SCAN name: myrac-scan.mydomain.com, Network: 1/192.xxx.xxx.0/255.255.255.0/nge1
SCAN VIP name: scan1, IP: /myrac-scan.mydomain.com/192.xxx.xxx.20
SCAN VIP name: scan2, IP: /myrac-scan.mydomain.com /192. xxx.xxx.21
SCAN VIP name: scan3, IP: /myrac-scan.mydomain.com /192. xxx.xxx.22
```

Die Clients erreichen die Cluster-Datenbanken nun ebenfalls über den SCAN-Eintrag in der tnsnames.ora oder aber den sonstigen Connect-Strings wie JDBC usw.
Die bisher verwendeten Host-Adresslisten sind also Geschichte. Ein Beispiel-tnsnames.ora-Eintrag für den Cluster-Datenbank-Service MALLPCL wäre also:

```
MALLPCL =
  (DESCRIPTION =
    (ADDRESS = (PROTOCOL = TCP)(HOST = myrac-scan.mydomain.com)(PORT = 1521))
    (CONNECT_DATA =
      (SERVER = DEDICATED)
      (SERVICE_NAME = MALLPCL)
    )
  )
```

Standby-Systeme

Jede der drei Cluster-Datenbanken bekommt eine physikalische Standby-Datenbank im Backup-Rechenzentrum (s. Abb. 1). Die Erstellung der Standby-Datenbanken erfolgte jeweils mittels Recovery Manager (RMAN) per DUPLICATE DATABASE-Kommando ohne Verwendung eines Backups direkt aus der Primär-Datenbank.

Die Standby-Datenbanken erhalten jeweils den *db_unique_name* der Primär-Datenbanken mit STB am Ende, z.B. MALLPSTB. Als *fal_server* dienen jeweils Services auf den RAC-Instanzen. Der Parameter *log_archive_dest_2* wird als Service zur Übertragung der archivierten Redolog-Dateien mittels Logwriter konfiguriert.

Da es sich bei dem Standby-System nicht um ein RAC, sondern einen Standalone-Server handelt, werden alle Cluster-Parameter vor Erstellung der jeweiligen Standby-Datenbank aus der init.ora bzw. dem spfile entnommen. Zusätzliche Parameter für den Data-Guard-Betrieb:

```
db_unique_name=MALLPSTB
*.log_archive_dest_1='LOCATION=/mnt/oracleasm/oraclearc/MALLP
VALID_FOR=(ALL_LOGFILES,ALL_ROLES) DB_UNIQUE_NAME=MALLPSTB'
*.log_archive_dest_2='SERVICE=MALLPCL LGWR ASYNC
VALID_FOR=(ONLINE_LOGFILES,PRIMARY_ROLE) DB_UNIQUE_NAME=MALLP'
*.log_archive_dest_state_1='enable'
*.log_archive_dest_state_2='enable'
*.log_archive_config='DG_CONFIG=(MALLPSTB,MALLP)'
fal_server='MALLP1_SRV','MALLP2_SRV'
*.standby_file_management='auto'
```

Im Fall eines Rollenwechsels darf nur eine Instanz der zu schwenkenden Cluster-Datenbank laufen. Die andere muss also ggfs. vor dem Rollenwechsel heruntergefahren werden.

Die Clients brauchen für den Fall des Rollenwechsels einen abweichenden tnsnames.ora-Eintrag, um den Standby-Server erreichen zu können. Dieser wird nicht vom SCAN abgedeckt.

Zusammenfassung

Wie immer in neuen Versionen muss man sich an Änderungen gewöhnen. Diejenigen, die die RAC-Installation und –Konfiguration betreffen, sind jedoch gut gelungen. Insbesondere die Verwendung nur noch eines einzigen Host-Namens für den Verbindungsaufbau der Clients ist in der Praxis hilfreich; auch im Hinblick auf die spätere Erweiterung eines Clusters, bei der dann keine Anpassungen der Client-tnsnames.ora mehr stattfinden müssen. Auch die Erstellung eines physikalischen Standby-Systems ist in Oracle 11g (allerdings bereits Release1) deutlich einfacher geworden (keine Umherkopiererei von Backups mehr). Lauffähigkeit und Performance des neuen Systems sind auf Seiten des Kunden absolut zufriedenstellend. Failover-Tests der redundant ausgelegten NetApp-Filer verliefen problemlos. Es lässt sich also festhalten, dass es nicht immer ASM sein muss, wenn es um den Aufbau eines RAC geht.

Kontaktadresse:

Susanne Jahr
Herrmann & Lenz Services GmbH
Höhestr. 37
D-51399 Burscheid

Telefon: +49 (0) 2174-6712-14
Fax: +49 (0) 2174-6712-22
E-Mail susanne.jahr@hl-services.de
Internet: www.hl-services.de