10gR2 Real Application Cluster leichtgemacht

Kunde:	DOAG*News*
Ort, Datum:	Artikel im Heft Q2 /2006
Thema / Themen:	**Artikel von merlin.zwo**
Projekt:	**10gR2 Real Application Cluster leichtgemacht**
Autor:	**Jochen Kutscheruk**

- Oracle & Technologien
- Systementwicklung
- Individuelle Lösungen

Einen Oracle Real Application Cluster selbst installieren? Alleine die Oracle-Dokumentation zu diesem Thema umfasst einige hundert Seiten. Welche Hardware braucht man? Welche Software? Wie funktioniert denn ein SAN?

Schon bei den Vorüberlegungen zu diesem Vorhaben kann man leicht die Übersicht verlieren. Und andererseits: die hausinterne Oracle-Datenbank läuft doch extrem zuverlässig und fällt nur ganz selten aus. (OK, dieses Argument ist tatsächlich nicht einfach von der Hand zu weisen). Andererseits - wenn etwas passiert, ist es meist sehr unangenehm und wirkt sehr lange nach.
Seit Oracle die RAC-Option ohne zusätzliche Lizenzkosten in der Standard Edition anbietet, ist die Hürde zur Entscheidung für einen RAC deutlich gesunken.
Dieser Artikel soll daher in kurzer Form eine gute, stabile und praxiserprobte Konfiguration sowie die notwendigen Voraussetzungen dafür beschreiben. Sie basiert auf normaler Intel-Hardware, (SuSE-)Linux Betriebssystem und einem SAN-Storage-System. Viele Hinweise aus der Oracle-Installationsdokumentation werden hier nicht aufgeführt, da sie auf einem SuSE-System bereits gegeben sind.

Die Voraussetzungen

- Zwei Server mit Single oder Dual Prozessor (EM64T durchaus empfehlenswert), 3GHz sind für gute Performance normalerweise mehr als ausreichend, 2 GB Hauptspeicher (es darf auch gerne mehr sein), zwei Gbit Netzwerkkarten, ein FibreChannel Hostadapter.
- SuSE SLES 9 (alternativ RedHat 3/4 AS/ES)
- SAN-Storage: hier gibt es zwischenzeitlich mehrere günstige und gute Anbieter; achten Sie darauf, daß durchgängig FibreChannel (bis hin zu den Festplatten) verwendet wird. SAN-Storages mit SATA-Festplatten funktionieren auch (und sind günstiger), allerdings zeigen diese bei hoher Last deutliche Performanceprobleme (diese Aussage bitte nicht verallgemeinern!).
- Ein kleiner Gbit-Switch für den Cluster Interconnect (die Verbindung zwischen den Servern), ein Crossover-Kabel funktioniert nicht wirklich!
- Und natürlich die Oracle 10g Software. Dieser Artikel bezieht sich auf die Version 10gR2. Verwenden Sie immer 32Bit Oracle-Software mit einem 32Bit Betriebssystem, die EM64T Oracle-Version für die entsprechende

Betriebssystem-Version. Die Installation einer Oracle 32Bit-Version auf einem EM64T-Linux-Betriebssystem wrd spätestens beim Linken scheitern.

Die Basis-Installation

- Bauen Sie die Netzwerkkarten und den FibreChannel-Hostadapter in die Servermaschinen ein. Stellen Sie alle Netzwerk- und FibreChannel- Verbindungen her. Konfigurieren Sie eine Partition auf dem SAN-System, auf die alle späteren RAC-Server parallel zugreifen dürfen. Diese Partition sollte im SAN-System über Raid1(0) oder Raid5 abgesichert sein. Einen relevanten Unterschied zwischen Raid5 und Raid1(0) konnten wir - allen gegenteiligen Aussagen zum Trotz - bei den heute verfügbaren schnellen SAN-Systemen nicht wirklich feststellen. Insofern bevorzugen wir eine Raid5-Konfiguration.
- Installieren Sie das SLES9-Betriebssystem mit ServicePack 2. Folgende "Selections" sollten Sie dabei mindestens auswählen:
  - Basis Runtime System
  - YaST
  - Graphical Base System
  - Linux Tools
  - KDE Desktop Environment (GNOME geht auch)
  - Authentication Server
  - C/C++ Compiler and Tools

  Es werden automatisch alle notwendigen Pakete in den passenden Versionen installiert.
  - Optional: orarun-Paket (SuSE-spezifisch) von der ServicePack 2 – CD

  Bei der Installation dieses letzten Pakets werden bereits viele notwendige Einstellungen vorgenommen (Oracle User, Kernelparameter,…). Im Service Pack 2 ist auch das ocfs2-Filesystem enthalten.
- Überprüfen Sie, daß die freigegebene Partition des der SAN-Storage erkannt wird: "fdisk -l" listet alle erkannten Festplatten auf. Die Partition auf der SAN-Storage ist als normale SCSI-Festplatte zu sehen (z.B. /dev/sdb).
- Konfigurieren Sie beide Netzwerkkarten. Die erste Netzwerkkarte erhält eine "normale" IP-Adresse aus Ihrem Netzwerk (z.B. 192.168.10.10 auf Server1, 192.168.10.11 auf Server2). Die zweite Netzwerkkarte erhält eine IP-Adresse aus einem nicht verwendeten (privaten) Nummernbereich (z.B. 172.16.10.10 auf Server1, 172.16.10.11 auf Server2). Dies ist die IP-Adresse für den Cluster Interconnect.
- Verbinden Sie die beiden ersten Netzwerkkarten mit Ihrem Netzwerkswitch, die Karten für den Cluster Interconnect mit dem kleinen Gbit-Switch (Sie können alternativ auch ein passendes VLAN auf einem gemanagten Switch einrichten).
- Reservieren Sie für jeden Server eine weitere IP-Adresse aus dem Adressbereich Ihres Netzwerks (z.B. 192.168.10.20 und 192.168.10.21). Dies sind die "virtuellen" IP-Adressen, über die später die Kommunikation mit den Clients läuft.

Alle beteiligten Server müssen zeitsynchron sein. Falls die Server direkten Kontakt zum Internet haben, erreichen Sie dies am einfachsten über den ntp-Daemon, der sich die aktuelle Zeit z.B. aus Braunschweig von den Servern ntp1.ptb.de und ntp2.ptb.de holt. Fügen Sie dazu die Zeilen

```
server ntp1.ptb.de
server ntp2.ptb.de
```

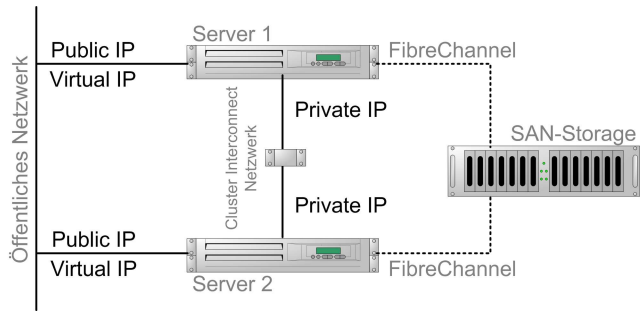
in die Datei /etc/ntp.conf ein und starten Sie den ntp-Daemon mit "rcxntpd start".

- Laden Sie das hangcheck-timer Modul:

```
/sbin/insmod hangcheck-timer hangcheck_tick=30 hangcheck_margin=180
```

Damit dieses Modul bei jedem Neustart geladen wird, fügen Sie diesen Aufruf in die Datei /etc/init.d/boot.local ein.

Erklärung der verschiedenen IP-Adressen



Für einen Oracle-Cluster werden 3 Gruppen von IP-Adressen benötigt:

1. ***Die Public IP:***
   Dies ist die ganz normale „öffentliche“ IP-Adresse des Servers („öffentliche“ bezieht sich dabei nicht auf einen speziellen IP-Adressraum, sondern nur auf die Tatsache, dass diese IP-Adresse aus dem lokalen Netzwerk kommt).
2. ***Die Private IP:***
   Über diese Verbindung tauschen die Clusterknoten intern Nachrichten und Daten aus. Dieses „Cluster Interconnect Netzwerk“ muß ein separater Adressbereich sein. Es darf keine Überschneidung mit dem „Öffentlichen Netzwerk (Public IP) geben!
3. ***Die Virtual IP:***
   Die Virtual-IP-Adressen des Clusters werden von den Clients für die Kommunikation mit der Datenbank verwendet. Der Client verbindet sich also nicht auf die Public IP, sondern ausschliesslich auf die Virtual IP. Auch der Datenbank-Listener lauscht nur auf die Virtual IP! Diese liegt im Netzwerk-Nummernbereich der Public IP.
   Der Grund für diese Konstruktion ist relativ einfach:
   Sobald ein Server ausfällt, wird dessen Virtual IP von einem verbleibenden Server übernommen. Wenn ein Client eine Verbindung zu dem ausgefallenen Server hatte, muß er nicht erst lange auf den TCP-Timeout warten, bevor die Verbindung

(TAF!) auf einen anderen Server gewechselt werden kann, er bekommt unmittelbar eine Antwort von dem Server, der die Virtual-IP übernommen hat.

Die Umgebung für die Oracle-Installation einrichten

- Legen Sie eine Gruppe "dba" an (z.B. über YaST).
- Legen Sie eine Gruppe "oinstall" an.
- Legen Sie den User "oracle" an. Dieser User sollte Mitglied der Gruppen "dba" und "oinstall" sein und /bin/bash als Shell haben. Die Default-Gruppe sollte "dba" sein. Weiteren Gruppen sollten dem "oracle"-User nicht zugeordnet sein.
- Die uid- und gid-Nummern müssen auf allen Servern identisch sein!
- Setzen der notwendigen Kernelparameter in /etc/sysctl.conf:

```
kernel.shmall           = 2097152
kernel.shmmax           = 2147483648 (physischer Speicher / 2)
kernel.shmmni           = 4096
kernel.sem        = 250 32000 100 128
fs.file-max             = 65536
net.ipv4.ip_local_port_range = 1024 65000
net.core.rmem_default   = 262144
net.core.rmem_max       = 1048576
net.core.wmem_default   = 262144
net.core.wmem_max       = 1048576
```

- Stellen Sie sicher, daß die Kernelparamter beim Booten gesetzt werden:

```
chkconfig boot.sysctl on
```

- Aktivieren Sie die geänderten Kernelparameter:

```
sysctl -p
```

- Erhöhen Sie die Shell-Limits für den User Oracle. Fügen Sie folgende Zeilen in die Datei /etc/security/limits.conf ein:

```
oracle      soft  nproc       2047
oracle      hard  nproc       16384
oracle      soft  nofile      1024
oracle      hard  nofile      65536
```

- Fügen Sie folgende Zeilen in /etc/profile.local ein, damit der User Oracle diese Limits bei jedem Anmelden gesetzt bekommt (erstellen Sie die Datei, wenn sie noch nicht vorhanden ist):

```
if [ "$USER" = "oracle" ]; then
      ulimit -u 16384 -n 65536
fi
```

- Überprüfen Sie die /etc/hosts-Datei: Es müssen folgende Einträge vorhanden sein:

```
127.0.0.1     localhost
192.168.10.10 server1.ihr-netz.de server1
192.168.10.11 server2.ihr-netz.de server2
192.168.10.20 server1-vip.ihr-netz.de server1-vip
192.168.10.21 server2-vip.ihr-netz.de server2-vip
172.16.10.10 server1-priv.ihr-netz.de server1-priv
172.16.10.11 server2-priv.ihr-netz.de server2-priv
```

Die Oracle-Clusterware auf den Servern ruft über ssh auf allen beteiligten Knoten Programme auf und führt Aktionen durch. Dabei darf von shh kein Passwort abgefragt werden, d.h. die Server müssen untereinander autorisiert sein, indem jeder Server den ssh-Schlüssel des anderen Servers kennt.

- Melden Sie sich als User oracle an
- Generieren Sie sich einen ssh-Key für Oracle:
  ```
  ssh-keygen -t rsa
  ```
- Kopieren Sie die erzeugte Datei $HOME/.ssh/id_rsa.pub in die Datei $HOME/.ssh/authorized_keys
- Wiederholen Sie diesen Vorgang auf jedem Server und kopieren Sie den Inhalt aller id_rsa.pub-Dateien in die obige authorized_keys-Datei.
- Kopieren Sie diese Datei auf jeden beteiligten Server in das .ssh-Verzeichnis des Users Oracle
- Ändern Sie als User root in der Datei /etc/ssh/ssh_config die Zeile
  ```
  StrictHostKeyChecking ask
  ```
  auf
  ```
  StrictHostKeyChecking no
  ```
  Dadurch wird die Sicherheitsabfrage beim Verbindungsaufbau zu noch unbekannten Servern unterdrückt.
- Testen Sie den problemlosen ssh-Zugang auf alle Server von allen Servern, z.B. mit "ssh server2 ls". Der Befehl sollte fehlerfrei und ohne weitere Rückfrage ausgeführt werden. Dies muß auch (sozusagen als loopback-Verbindung) auf den eigenen Server funktionieren!
- Überprüfen Sie, daß die Umgebungsvariablen LANG, NLS_LANG, ORACLE_HOME und ORACLE_BASE nicht gesetzt sind. Speziell mit der Default-Einstellung für LANG (de_DE.UTF-8) kann der Installer Probleme bekommen.
- Entpacken Sie die Installations-CD in ein Verzeichnis, z.B. /tmp/oinstall
- Führen Sie als User root auf jedem Server das rootpre.sh-Skript aus (nur auf x86_64-Systemen).
  Wechseln Sie dazu in das Verzeichnis /tmp/oinstall/clusterware/rootpre und rufen Sie
  ```
  ../rootpre.sh auf.
  ```

Letzte Vorbereitungen

Nachdem alle diese Voraussetzungen geschaffen sind, ist die eigentliche Cluster-Installation ein Kinderspiel.

Zuerst müssen Sie noch festlegen, mit welcher Storage Option Sie die Datenbank installieren möchten. Bei der Standard Edition wird Ihnen die Wahl leicht gemacht, hier geht ausschliesslich Automatic Storage Management (ASM). Bei der Enterprise Edition haben Sie die Wahl:

Storage Option	OCR und Voting Disk	Oracle Software	Datenbank	Recovery
ASM	Nein	Nein	Ja	Ja
OCFS2	Ja	Ja	Ja	Ja
NFS	Ja	Ja	Ja	Ja
RAW	Ja	Nein	Ja	Nein

Dies bedeutet, daß Sie bei der Oracle Standard Edition für die Oracle Cluster Registry (OCR) und die Voting Disk extra eine OCFS2-Partition anlegen müssen. Alternativ wäre auch ein RAW-Device möglich. Diese Daten auf ein NFS-Share zu legen ist leider keine Option, da dies tatsächlich nur mit zertifizierten NAS-Storages funktioniert.

Die OCR und Voting Disk können ab Oracle 10gR2 auch gespiegelt werden - diese Option sollte man unbedingt nutzen, sofern man unabhängige Plattensubsysteme zur Verfügung hat.

Das einfachste Layout für die Oracle-Partitionen auf der SAN-Storage sähe dann folgendermaßen aus:

Partition	Größe	Inhalt	Storage Option
/dev/sd?1	ca. 120 MB	OCR und Voting Disk	OCFS2
/dev/sd?2	Ausreichend Platz für Tablespace-Files	Datenbank	ASM
/dev/sd?3	Ausreichend Platz für Archive Logs	Archive Logs bzw. Flash Recovery Area	ASM

Dies ist nicht die optimale Verteilung - eine etwas großzügigere Aufteilung auf verschiedene Platten wäre der Sicherheit keinesfalls abträglich.
Noch ein Hinweis zu LVM: Die Verwaltung der Platten wird durch LVM deutlich vereinfacht. Aber: LVM ist nicht "Cluster-Aware" und daher für unsere Zwecke nicht zu gebrauchen!
Erstellen Sie auf der ersten Partition ein ocfs2-Filesystem und mounten Sie dieses (z.B. nach /opt/oracle/oracrs).
Das komplette /opt/oracle-Verzeichnis sollte dabei dem User oracle gehören.

Installation der Oracle Clusterware

Überprüfen Sie zuerst die korrekte Konfiguration mit Hilfe der Cluster Verification Utility (CVU):

```
/tmp/oinstall/crs/Disk1/cluvfy/runcluvfy.sh stage -pre crsinst -n
server1,server2
```

Dieses Utility überprüft, ob alle notwendigen Voraussetzungen für die Installation erfüllt sind:
Node Reachability, User Equivalence, Node Connectivity, Administrative Privileges, Shared Storage Accessibility, System Requirements, Kernel Packages, Node Applications.
Alle Überprüfungen sollten fehlerfrei laufen. Eventuell kann bei der Virtual IP - Überprüfung noch ein Fehler auftauchen: entweder moniert das Tool, daß die Virtual IP noch gar nicht eingerichtet ist, oder daß die Virtual IP bereits vergeben ist (da Sie diese IP eventuell bereits an ein Interface gebunden haben). Im zweiten Fall sollten Sie die IP wieder vom Interface entfernen, den ersten Fall können Sie getrost ignorieren.

Die eigentliche Installation

Nachdem alle diese Voraussetzungen geschaffen wurden, ist die eigentliche Installation tatsächlich einfach. Sie funktioniert (fast) nach dem üblichen Schema einer Datenbank-Installation.

Installieren Sie die Cluster Ready Services

Rufen Sie dazu als User Oracle im Clusterware-Unterverzeichnis des Installationsverzeichnisses den runInstaller auf.
Bei der Abfrage der Clusterkonfiguration achten Sie darauf, daß der Public Node Name, Private Node Name und Virtual Host Name korrekt gesetzt sind. Folgen Sie ansonsten den Angaben des Installationsassistenten.
Der Speicherort für die Oracle Cluster Registry wäre z.B. /opt/oracle/oracrs/ClusterRegistry, für die Voting Disk /opt/oracle/oracrs/VotingDisk (im Verzeichnis /opt/oracle/oracrs befindet sich das vorhin gemountete ocfs2-Filesystem).

In unserer einfachen Konfiguration wählen Sie jeweils "external Redundancy" aus. Über "normal Redundancy" könnten Sie die OCR und Voting Disk auf verschiedene Platten spiegeln.
Die Installation sollte normalerweise fehlerfrei durchgeführt werden. Nach diesem Schritt haben Sie das Schwierigste auf jeden Fall hinter sich.

Installieren Sie die Oracle-Datenbank
Diese Installation funktioniert tatsächlich (fast) analog zu einer „normalen“ Datenbankinstallation.
Sie müssen bei der Installation lediglich zusätzlich angeben, auf welchen Knoten die Oracle-Software installiert werden soll. Die Installation wird von Ihnen selbst nur auf einem Knoten durchgeführt, die Cluster Services sorgen dafür, daß die Software automatisch auf alle anderen beteiligten Knoten verteilt und passend konfiguriert wird.
Bei der Installation der Datenbank können Sie direkt die „Automatic Storage Management“ Option anwählen. Damit wird die notwendige ASM-Instanz (für die Oracle Standard Edition) automatisch vor dem Anlegen der Datenbank eingerichtet.

Abschliessend
Sie sollten auf jeden Fall nach der Installation die Oracle Cluster Registry und die Voting Disk sichern. Diese Aktion sollte auch bei jeder Veränderung des Clusters (z.B. Knoten hinzufügen oder entfernen) durchgeführt werden.
Die Datenbank selbst sollte nicht mehr über *sqlplus*, sondern nur noch über *srvctl* gestartet und gestoppt werden. Das Starten und Stoppen der Datenbank wird (unter anderem) durch dieses Utility auf allen Knoten des Clusters durchgeführt.

Und schliesslich: Obwohl Sie mit dieser Kurzanleitung einen Oracle Real Application Cluster aufsetzen können, wird Ihnen trotz allem das Studium der entsprechenden Handbücher nicht erspart bleiben. Die ganzen Möglichkeiten, die Ihnen ein Cluster bietet, aber auch das notwendige KnowHow für den täglichen Betrieb, kann man leider nicht in einem kurzen Artikel vermitteln.

Kontakt:
Jochen Kutscheruk
jochen.kutscheruk@merlin-zwo.de