Einfach erklärt:
RAC Grundlagen für Dummies

Ralf Appelbaum
TEAM GmbH
Paderborn

Schlüsselworte
Real Application Clusters, RAC, Grid Infrastructure, Automatic Storage Management, ASM, ASM Cloud File System, ACFS, Oracle Cloud File System, CloudFS, Oracle Datenbank, RAC One Node, Administration, Installation, Single Client Access Name, SCAN, Oracle Cluster Registry, OCR, Oracle Local Registry, OLR, Votingdateien, Rolling Update, Flex ASM, Application Continuity, Maximum Availability Architecture, MAA,

Einleitung
Mit meinem Vortrag möchte ich die technischen Grundlagen und Begriffe eines Oracle Real Application Clusters (RAC) den unbewanderten Administratoren, Systemarchitekten und Entscheidern allgemeinverständlich bzw. illustrativ näher bringen.

Dabei werden die wesentlichen Komponenten des gesamten RAC Stacks, wie z.B.

- Oracle Clusterware,
- Grid Infrastructure,
- Automatic Storage Management (ASM),
- ASM Cloud File System (ACFS),

mit den Begriffen und Abkürzungen genannt, die Zusammenhänge dargestellt und auch einige Hinweise zu Best-Practice gegeben.

Die neuen Features der Version 12c, wie z.B. Flex ASM und Application Continuity, werden dargestellt und selbstverständlich werde ich auch die wichtigsten Links im Netz liefern.

Ich wünsche mir, dass am Ende des Vortrags

- Entscheider die Begriffe und Abkürzungen im RAC Umfeld, die sie in Projekten immer wieder zu hören bekommen, zuordnen können,
- Systemarchitekten die strukturellen Voraussetzungen eines RAC kennen und im Groben berücksichtigen lernen,
- Administratoren einfache Anleitungen im Netz zur Installation eines Test-/Spiel-RAC nachvollziehen können.

Was ist RAC? Wozu RAC?
Es gibt zwei wesentliche Gründe, einen Oracle **R**eal **A**pplication **C**lusters (RAC) einzusetzen:

- Hochverfügbarkeit
- Skalierbarkeit

Oracle RAC ist ein Lösungsbaustein im Bereich Hochverfügbarkeit von Oracle **R**elationalen-**D**atenbank-**M**anagement-**S**ystemen (RDBMS). Ein Oracle RDBMS gliedert sich in zwei Grundkomponenten:

Datenbank,	eine Menge von Dateien welche die Daten persistent speichern
Instanz,	die Laufzeitumgebung bestehend aus Hauptspeicher und Prozessen welche den Zugriff auf die Daten ermöglichen

Die Datenbank bzw. Dateien werden in der Regel abgelegt auf einem zentralen Storage:

Storage **A**rea **N**etwork (SAN)

Network **A**ttached **S**torage (NAS)

Diese stellen die Verfügbarkeit der Dateien über redundante, d.h. mehrfache Speicherung, in diversen RAID (**R**edundant **A**rray of **I**ndependent **D**isks) Leveln sicher.

Der verbleibende **S**ingle **P**oint **o**f **F**ailure (SPOF) des Oracle RDBMS, welcher bei Ausfall also einen Totalausfall des RDBMS bewirkt ist die Instanz.

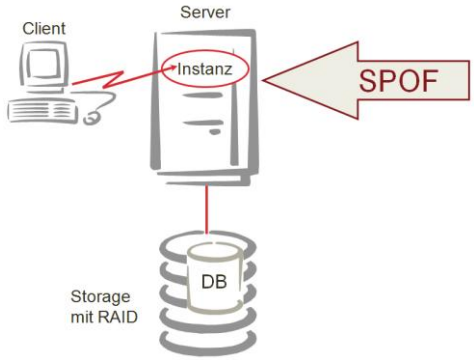

Abb. 1: Single Point of Failure (SPOF)

Bei einer RAC Konfiguration werden zwei oder mehr Instanzen auf jeweils eigenen Servern gegen eine gemeinsame Datenbank auf einem zentralen Storage konfiguriert. Für den Client bildet diese Konfiguration eine Black Box. Der Zugriff auf das **D**atenbank-**M**anagement-**S**ystemen (DBMS) erfolgt weiterhin über SQL*Net mittels „Servername:Port/Datenbankservice“ wobei der Servername durch einen **S**ingle **C**lient **A**ccess **N**ame (SCAN) ersetzt wird, welcher als DNS-Name über ein **D**omain **N**ame **S**ystem (DNS) auf 1 bis 3 IPs im Round-Robin-Verfahren aufgelöst wird.

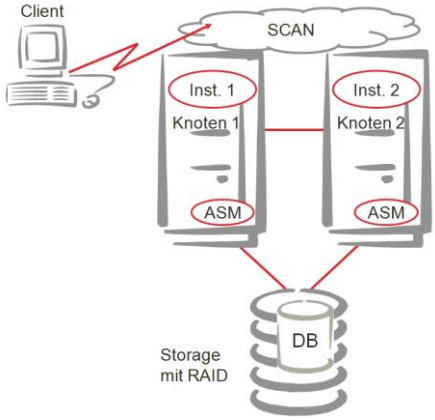

Abb. 2: Oracle Real Application Clusters (RAC)

Mit der Redundanz der Serverhardware ergibt sich der zweite Aspekt eines RAC, die horizontale Skalierbarkeit. Mit jedem Knoten kommen weitere CPU Ressourcen, weiterer Hauptspeicher, weiterer I/O- und Netzwerk-Durchsatz hinzu.

Hochverfügbarkeit mit RAC

Ein RAC alleine gewährleistet noch keine vollständige Hochverfügbarkeit. Dem Ausfall z.B. eines kompletten Storage, der Stromversorgung, der Netzwerkanbindung oder gar des ganzen Rechenzentrums (RZ) entgeht man damit nicht. Dem so genannten Disaster-Fall begegnet man, indem das komplette Oracle DBMS in einem anderen RZ bzw. einer anderen Brandschutzzone bzw. an

einem weiteren Standort redundant vorgehalten wird. Die Daten werden dabei z.B. mittels Oracle Data Guard vom primären DBMS auf ein oder mehrere sekundäre DBMS gespiegelt.

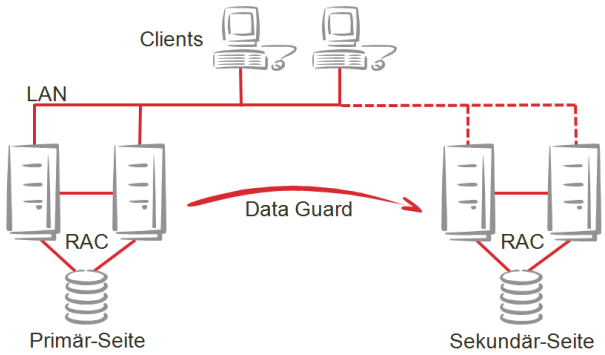

Abb. 3: Maximum Availability Architecture (MAA)

Zusammen in einer Konfiguration bilden RAC und Data Guard die **M**aximum **A**vailability **A**rchitecture (MAA).

Eine beschränkte alternative zur MAA stellt ein RAC auf einem Extended Distance (Stretched) Cluster dar, besonders beliebt bei Standard Edition mit welcher Data Guard nicht genutzt werden darf.

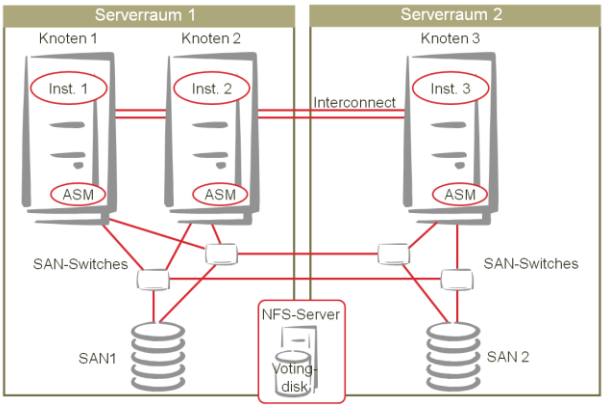

Abb. 4: Extended Distance (Stretched) RAC

Dabei werden die Knoten eines RAC auf zwei (bis drei) RZ‘s bzw. Brandschutzzonen mit jeweils eigenem SAN verteilt. Größere Distanzen zwischen den redundanten Teilen sind jedoch auf Grund des negativen Einflusses der Latenz auf die Performance und RAC Funktionalität nicht möglich.

RAC Komponenten

Der Oracle RAC nutzt die Oracle Grid Infrastruktur (GI). Diese bildet den Hochverfügbarkeits Framework nicht nur für RAC sondern auch für andere Oracle Software Produkte. Die GI beinhaltet folgende Komponenten:

- Oracle Clusterware (OCW),
- Oracle Automatic Storage Management (ASM),
- ASM Cloud File System (ACFS) bzw. Oracle Cloud File System (CloudFS),
- Single Client Access Name (SCAN),
- SCAN Listener,
- Grid Naming Service (GNS).

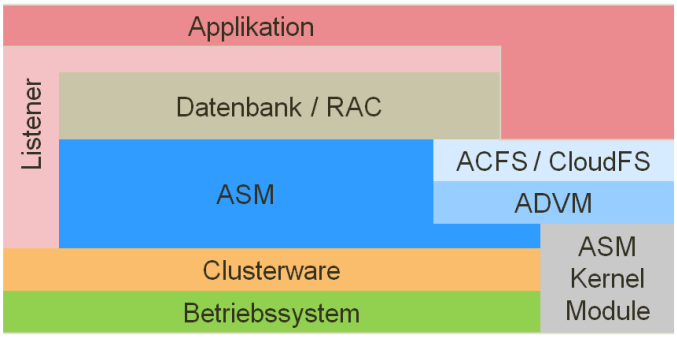

Abb. 5: RAC / GI Softwarestack

Die **O**racle **C**luster**w**are (OCW) ist integraler Bestandteil der GI und macht aus einer Menge einzelner Server einen Cluster. Die Clusterware verwaltet und überwacht Ressourcen im Cluster, verteilt sie über die Knoten im Cluster und startet sie ggf. neu und vieles mehr.

Die Datenbank, d.h. die Dateien des DBMS, muss im RAC auf einem zentralen Storage liegen und von allen Instanzen gleichzeitig les- und schreibbar sein. Oracle empfiehlt hierfür das **A**utomatic **S**torage **M**anagement (ASM), das bei RAC mit Standard Edition auch verpflichtend ist.
ASM erwartet einen „Bulk of Disks“, verwaltet die Platten wie ein Volumemanager, den es auch ersetzt, realisiert Redundanz über Spiegelung und stellt damit das Dateisystem für die Datenbank. Die Datendateien werden im ASM in Plattengruppen gespeichert, die abhängig von der Redundanz aus einer unterschiedlichen Zahl an Fehlergruppen bestehen müssen.

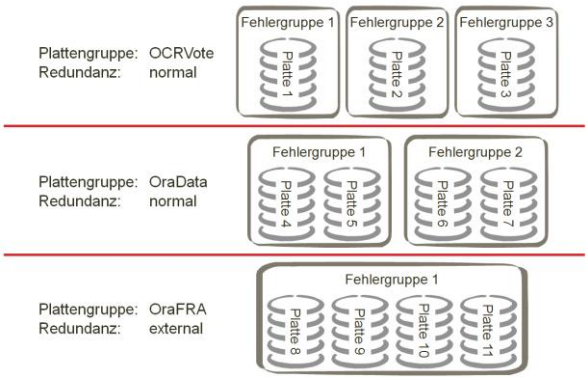

Abb. 6: ASM Plattengruppen und Redundanzen

Es gibt die drei folgenden Redundanz-Level mit folgenden Eigenschaften:

Redundanz Typ	Standard Spiegelung	minimale Anzahl Fehlergruppen	tollerierbarer Ausfall
externe Redundanz	keine / ungesichert	1	keiner
normale Redundanz	zwei-Wege	2	1 FG
hohe Redundanz	drei-Wege	3	2 FG

Das ASM verwaltet nur die Metadaten für die Inhalte der Plattengruppen. Das DBMS führt das Lesen und Schreiben von Oracle Blöcken über die Instanzprozesse selber aus. Dabei wird das Betriebssystem I/O nicht genutzt. Die Plattengruppen sind auch im Betriebssystem nicht als Dateisystem sichtbar bzw. die Datenbankdateien sind aus dem Betriebssystem nicht direkt erreichbar.

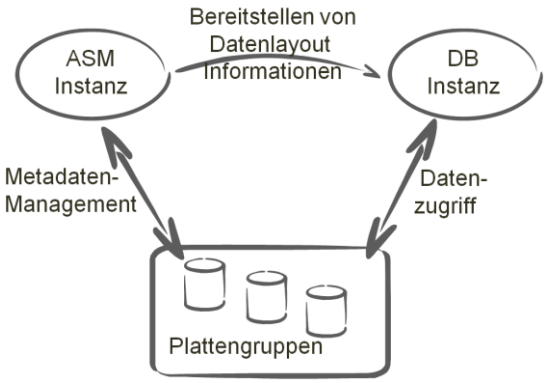

Abb. 1: Relation zw. ASM Instanz und Datenbank Instanz

Bis Oracle Version 11g muss auf jedem Knoten im RAC eine ASM Instanz laufen, die von den Prozessen vergleichbar einer Datenbank Instanz ist. Ab Oracle 12c gibt es das so genannte Flex ASM als alternative Installationsoption. Dabei muss nicht mehr auf jedem Knoten eine ASM Instanz laufen, sondern es könnten beispielsweise bei einem 6-Knoten Cluster nur noch 3 ASM Instanzen aktiv sein.

Rolling Updates und Patches
Ein RAC ermöglicht es, die Knoten einzeln, nacheinander mit Updates und Patches für Betriebssystem, Oracle Software oder anderes zu versorgen.

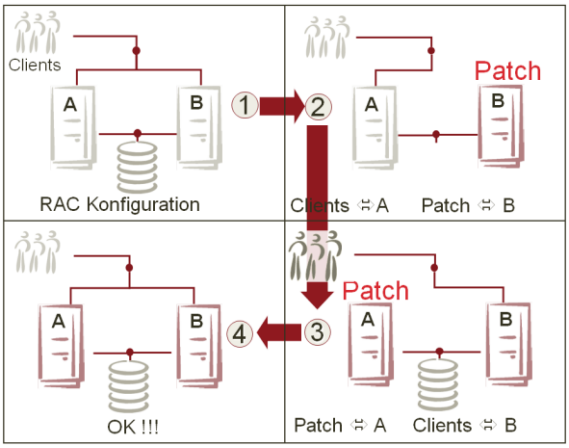

Abb. 1: Rolling Update im RAC

Die Datenbank bleibt dabei für die Applikation permanent verfügbar.

Netzwerkkonfiguration
Die Netzwerkvoraussetzungen für einen Oracle Cluster sind ein wichtiger Teil der Installation. Für einen Cluster werden eine ganze Reihe IP's benötigt:

- Pro Server 1 öffentliche IP und 1 Hostname
- Pro Server 1 virtuelle IP und 1 weiterer Hostname (im selben Netzwerk wie die öffentliche)
- Pro Server 1 IP für Interconnect (privates Netzwerk)
- Für den gesamten Cluster 1-3 IP's (im selben Netzwerk wie die öffentliche)
  Diese 3 IP's werden im Round-Robbin Verfahren unter EINEM Hostnamen (bis 15 Zeichen lang) referenziert

Beispiel:

Hostname	IP	virt.IP	virt. Name	Interconnect	Interconnect-Name
rac1	10.0.0.1	10.0.0.11	rac1-vip	192.168.0.1	rac1-priv
rac2	10.0.0.2	10.0.0.12	rac2-vip	192.168.0.2	rac2-priv

Zusätzlich die 1-3 clusterweiten IP-Adressen:

Hostname	IP-Adressen
rac	10.0.0.21, 10.0.0.22, 10.0.0.23

D.h. je eine der öffentlichen und je eine der privaten IP's werden bereits bei der Linux-Installation den Netzwerkkarten zugeordnet. Während der Oracle-Installation benötige man dann nochmal IP's aus dem öffentlichen Bereich für die VIP's der einzelnen Server und 1 bis 3 IP's aus dem öffentlichen Bereich für den gesamten Cluster.

Für beide Netze (öffentlich und Interconnect) sollten jeweils zwei physikalische Netzwerkports verwendet werden. Die zwei Netzwerkports im öffentlichen Netz müssen mittels Bonding (Failover) zu einem logischen Netzwerkport konfiguriert werden. Die Netzwerkports im privaten Netz (Interconnect) können ebenfalls mittels Bonding (Failover) zu einem logischen Netzwerkport konfiguriert werden, das kann aber auch Oracle übernehmen und dann ein Loadbalance und Failover darüber selber realisieren (Letzteres ist von Oracle empfohlen). Für den Interconnect bzw. auf dessen Switch muss Multicast aktiviert sein.

Auf den Knoten sollte Network Time Protocol (NTP) entweder vollständig konfiguriert und aktiv sein oder gar nicht konfiguriert sein. Nicht konfiguriert bedeutet, es darf keine ntp.conf Datei vorhanden sein. Ist kein NTP konfiguriert, dann richtet die Oracle Grid Infrastruktur einen eigenen Dienst ein um die Zeiten der Knoten identisch zu halten.

Die hier aufgeführten Punkte, aber auch noch eine Reihe weiterer, die sich u.a. aus den oben aufgeführten Schlüsselworten ergeben, werden im Vortrag erläutert.

Kontaktadresse:
Ralf Appelbaum
TEAM GmbH
Hermann-Löns-Str. 88
D-33104 Paderborn

Telefon: +49 (0)5254 / 8008-37
Fax: +49 (0)5254 / 8008-19
E-Mail: ra@team-pb.de
Internet: http://www.team-pb.de