



RAC auf Sun Cluster 3.0

Schlüsselworte

RAC, OPS, Sun Cluster, Performance, Availability

Zusammenfassung

Oracle hat mit dem Real Application Cluster (RAC) aus einer Hochverfügbarkeitslösung eine Höchstverfügbarkeitslösung geschaffen, auch bekannt als "The Unbreakable Database System". Wieviel Performance und wieviel Availability damit zur Verfügung stehen, wird dargestellt.

HA für Oracle – Ein Vergleich

Ein Vergleich der Lösungen Data Guard, HA Oracle, RAC ist nicht leicht, weil die Ansätze verschieden sind und nicht zum gleichen bzw. gewünschten Ziel führen. Zum Vergleich werden 6 Kriterien herangezogen.

- Wieviel HA darf es denn sein ?
- Wie lange dauert ein Failover ?
- Ist im Failover Fall mit Datenverlust zu rechnen ?
- Was sagt die Performance dazu ?
- Ist die Lösung gut und einfach wartbar ?
- Wieviel kostet die Lösung ?

Folgende Tabelle soll als erste Entscheidungshilfe dienen:

Criteria	Data Guard	HA Oracle	RAC
<i>Availability</i>	Normal	High	Very High
<i>Failover</i>	8-10 min	2-3 min	< 1 min
<i>Data Loss</i>	High	Low	Very Low
<i>Performance</i>	Low	Normal	High
<i>Manageability</i>	Complex	Easy	Easy
<i>Cost</i>	Low	Normal	High

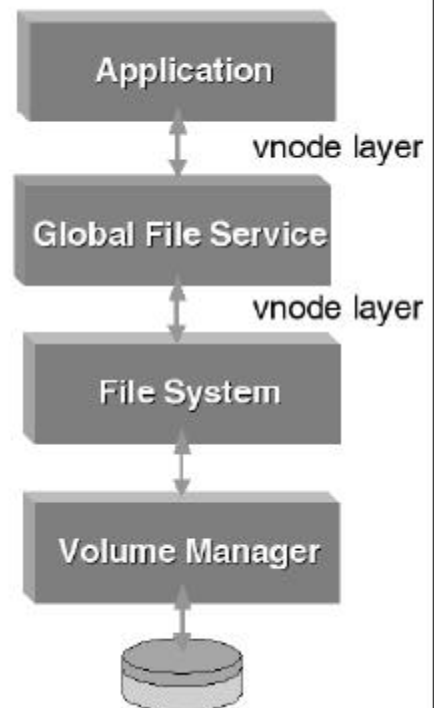
Sun Cluster 3.0 - Key Features

Wesentliche Features des 3.0er Release sind die Integration des Clusters im Solaris 8 Kernel, das clusterweite Filesystem und das Built-in Load Balancing. Letzteres wird leider vom RAC nicht genutzt, so dass nicht über alle Interconnects hinweg skaliert wird, selbst der "cluster_interconnects" Parameter hilft nicht wirklich.

Was bei der Administration hilfreich ist, ist das clusterweite Filesystem, auch Global FS genannt. Es ist ein hochverfügbares, verteiltes und cache-coherentes FS. Mit nur einem einzigen Mount Befehl (mount -o global) ist das Global FS auf allen Knoten transparent. Hier können Oracle Binaries und Konfigurationen für alle Knoten global abgelegt werden.

Global File Service

- Kein neues FS
- Cluster File System
 - hochverfügbar, distributed, cache-coherent
- Kernel-basierte Client/Server Architektur
 - PxFs Mechanism basiert auf *vnode* interface
- unabhängig vom FS Type und Volume Manager
- Failover/Switchover transparent zur Application und zum User
- Global Mount von einem Knoten für alle Knoten
 - mount -o global
 - /etc/vfstab



Der Cluster bootet automatisch im Cluster Mode. Updates und Patches können im Non-Cluster-Mode Knoten für Knoten installiert werden, ohne den gesamten Cluster zu stoppen.

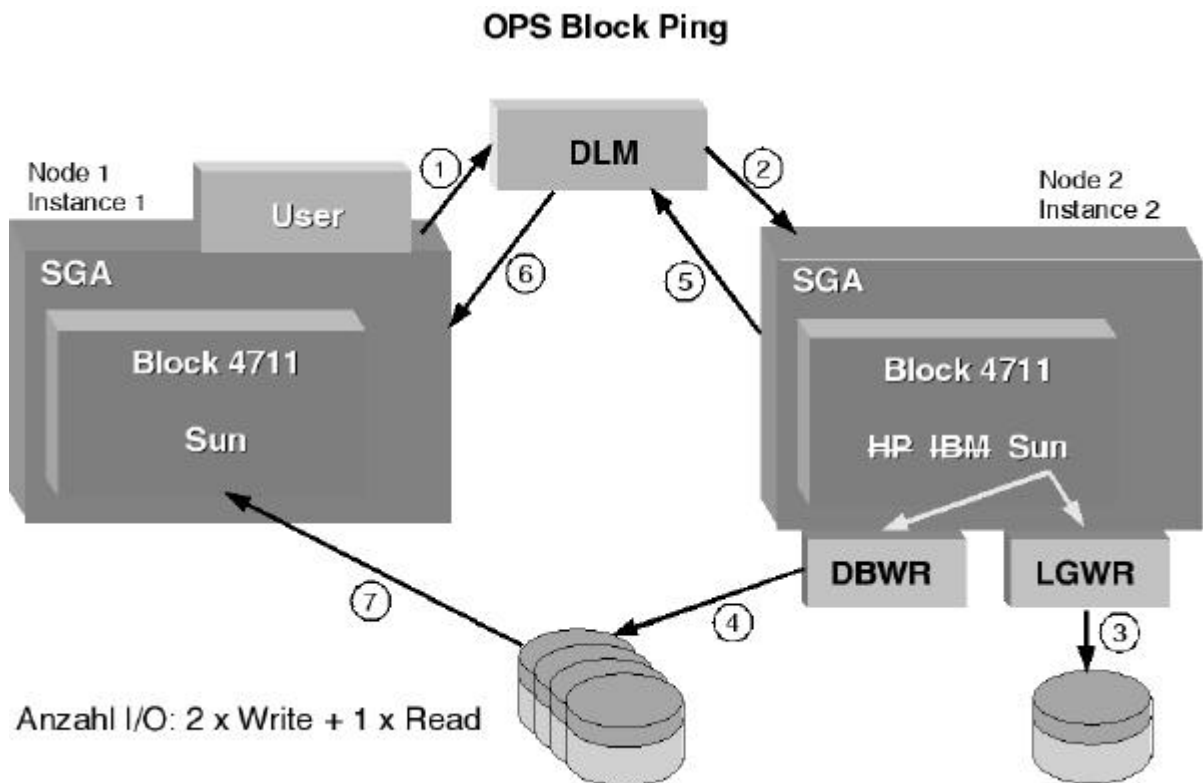
Während der Real Application Cluster sich um seine User Connects kümmert, kümmert sich der darunterliegende Sun Cluster um Hardware Fehler.

Von OPS nach RAC auf Sun Cluster

Beim RAC ist es gelungen das Cache Coherency Protokoll zu optimieren. Dabei galt es, die Anzahl Messages zu reduzieren und I/Os zu vermeiden. Durch Cache Fusion entsteht der Eindruck einer globalen SGA über alle Knoten.

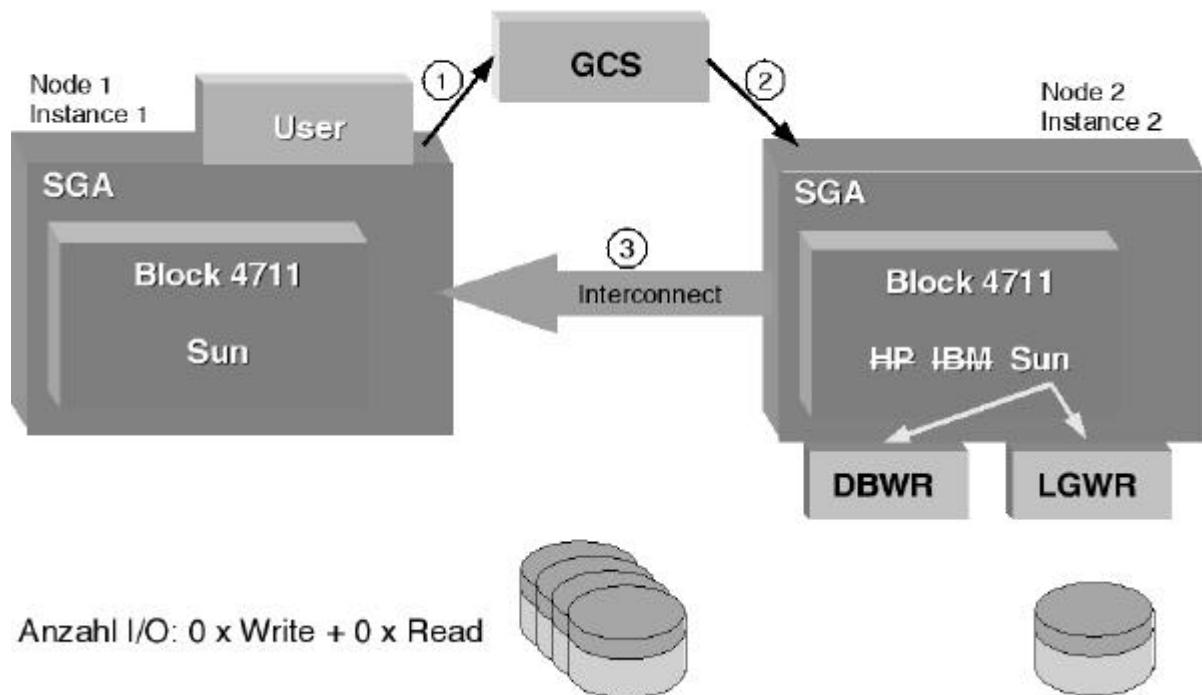
Was macht den RAC schneller als den OPS ?

Block Shipping statt Block Ping.



Während beim OPS der Block Ping und damit 2 Write und 1 Read I/O für die Schreibkonsistenz nötig war, ist beim RAC kein Disk I/O mehr nötig. Die Daten werden von SGA zu SGA über den Cluster Interconnect geschoben. Die Anzahl Messages wurde reduziert und I/Os vermieden. Der Cache bleibt ohne I/O coherent.

RAC Block Shipping



Laut Oracle skaliert der RAC über Rechengrenzen hinweg fast linear. Beim 2 Knoten Cluster skaliert er 95% und beim 4 Knoten Cluster 89%. Die Skalierung hängt jedoch stark von der Applikation ab, je nach dem, wie die Tabellen abgefragt werden. Mehr dazu folgt im nächsten Abschnitt.

Eine weitere Neuerung ist das Connect Load Balancing. Die Listener der Instanzen tauschen sich alle 30 Sekunden über den Node, Instanz und Dispatcher Load aus. Beim Connect wird vom Listener entschieden auf welche Instance der Reconnect geht. So landet man stets auf der am wenigsten ausgelasteten Instanz.

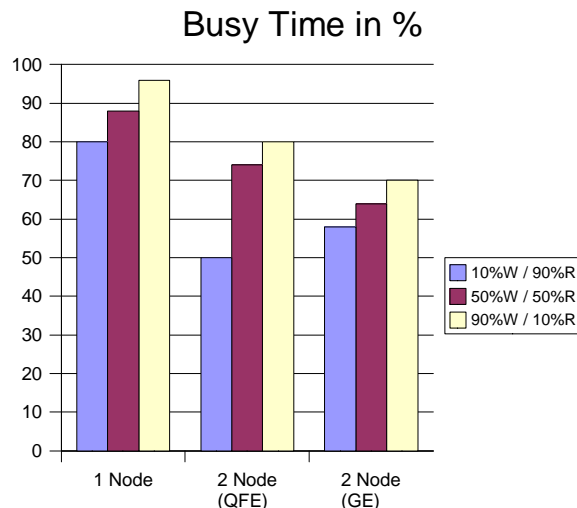
“best practice” Konfiguration

Availability und Performance sind der Wunsch an die Datenbank. Wie kann beides erreicht werden ?

- Availability wird durch horizontale Skalierung erreicht
- Performance wird durch vertikale Skalierung erreicht

Viele kleine Server produzieren Overhead. Die Folge ist, daß die Systemlast aller Server steigt. Eine Studie zeigt, daß bei einem Server die Systemlast bei 80% liegt und das Hinzufügen eines weiteren Servers die Systemlast zu 50% auf den einen und 50% auf den anderen Server verteilt wurde. Das ergibt einen Overhead von 20% bei Lastverteilung.

Es empfiehlt sich, besser wenige große Server im RAC zu nutzen als viele kleine Server.



Transparent Application Failover (TAF) ist das Feature, bei dem die Connects automatisch und transparent zur nächsten laufenden Instanz geswitched werden. Beim Herunterfahren (shutdown immediate) oder Absturz (shutdown abort) einer Instanz funktioniert der Reconnect. Der Failover dauert weniger als eine Minute. Beim Power Off Failing eines Knotens bekommt der Client zunächst nichts vom Ausfall mit. Erst nachdem TCP/IP Timeouts abgelaufen sind, wird der Ausfall bekannt und der Failover eingeleitet. Der Failover beim Power Off dauert gemäß Solaris TCP/IP Defaults 11 Minuten. Die Thresholds, die zu setzen sind, um dieses Problem zu umgehen, sind:

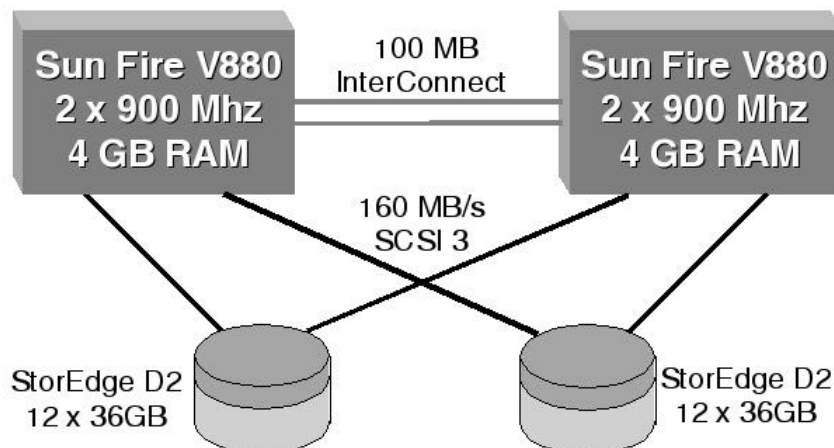
- tcp_ip_abort_interval (default: 480000ms)
- tcp_ip_abort_cinterval (default: 180000ms)

Die Defaultwerte sind natürlich abhängig vom Betriebssystem !

Das Projekt "Wirecard"

Bei diesem Projekt standen zwei Lösungen auf dem Plan, entweder HA Oracle oder RAC. HA Oracle läuft im Sun Cluster mit einer oder mehr Instanzen, die automatisch im Fehlerfall zum noch lebenden Knoten migrieren. Ein Aktiv / Aktiv Konzept ist auch möglich. Großer Nachteil ist, daß beim Failover nicht abgeschlossene Transaktionen verloren gehen und deswegen Clients ihre Abfragen erneut absetzen müssen. Möglichst keine Transaktionen zu verlieren wurde ein zunehmendes Kriterium. Somit kam als endgültige Lösung nur RAC in Frage.

Projekt WireCard



Ein 2 Knoten Cluster mit 2 SunFire V880 inklusiv 2 x 900MHz und 2 GB RAM waren für die Hochverfügbarkeit ausreichend. Da die Wachstumsrate der User eine Unbekannte war, wurde auf die horizontale Skalierung für die Performance geachtet. Die SunFire V880 kann noch auf maximal 8 CPUs und 32GB RAM erweitert werden.

Fazit

Oracle hat wahrhaftig mit den RAC "The Unbreakable Database System" geschaffen. Logische oder Benutzer Fehler werden nicht toleriert. Die darunterliegende Cluster Software sollte Hardware Fehler wie Netzwerk- oder Plattenpfade abfangen können. Ein Betriebssystem mit "Dynamic Reconfiguration" verringert die Downtime.

Welche HA Lösung für Oracle die ideale ist, hängt von der Anforderung ab. Für jede Anforderung gibt es ein passendes Produkt.

Kontakt:

Marco Kühn

best Systeme GmbH
Münchener Strasse 123a
85774 Unterföhring

Tel.: 089-9506080
Fax: 089-9506070
kuehn@best.de