Oracle Grid Infrastructure 11gR2 im Einsatz

Jochen Kutscheruk
merlin.zwo InfoDesign GmbH & Co. KG
Karlsruhe

Schlüsselworte:

Grid Infrastructure, RAC, Cluster, SCAN, ACFS, GNS, Policy Based Cluster

Einleitung

Eigentlich hätte man bei der Oracle Grid Infrastructure 11gR2 eine moderate Weiterentwicklung der Version 11gR1 erwartet. Bei näherer Betrachtung entpuppt sich die Grid Infrastructure jedoch als grundlegend überarbeitetes Konzept. Inwieweit dieses neue Konzept sinnvoll ist und was es beim praktischen Einsatz zu beachten gilt soll in diesem Vortrag betrachtet werden.

Verbesserungen bei der Installation

Einige Dinge wurden beim Oracle Universal Installer drastisch verbessert.

SSH Setup im Installer integriert:
Gerade das SSH Setup (Authentifizierung über Schlüsselaustausch) hat bei der Installation eines Real Application Clusters einiges an Aufwand verursacht. Immer wieder geschah es, dass man – gerade bei vielen beteiligten Knoten – doch einen Knoten vergessen oder nicht alle Verbindungsmöglichkeiten durchgetestet hatte. Dies wird jetzt durch das integrierte SSH Setup bei der Installation „auf Knopfdruck“ erledigt.

Die Cluster Verification Utility im Installer integriert:
Auch hierdurch wird die Fehlerquote bei der Installation deutlich verringert. Während der Installation wird nochmals ein umfassender Check durchgeführt, ob alles korrekt ist.

Erzeugung automatischer Fixup-Scripte vom Installer:
Sollte der Installer bei der Installation einen Fehler feststellen – z.B. falsche Kernelparameter – so wird automatisch ein Fixup-Script erzeugt, mit dem man dieses Problem beseitigen kann. Auch dies erleichtert die Installation erheblich.

Was gibt es an neuen Konzepten?

Es gibt eine ganze Zahl neuer Konzepte, angefangen beim RAC Single Node, von denen in diesem Vortrag folgende betrachtet werden sollen:

- Integration von ASM in die Grid Infrastructure
- Cluster Registry und Voting Disk im ASM
- ASM Cluster File System
- Redundant Interconnect
- SCAN als Erweiterung zur VIP
- Grid Naming Service

- Cluster Time Synchronization Service
- Policy Based Cluster

Integration von ASM in die Grid Infrastructure

Das ASM gehört jetzt nicht mehr zum Datenbank-Installationsbereich, sondern wird bei der Grid Infrastructure mit installiert und konfiguriert. Dieser Schritt ist konsequent und logisch, da die Grid Infrastructure nach dem Willen von Oracle zu mehr dienen soll als nur der Ablage von Datenbankdateien (siehe auch ASM Cluster File System). Daher gehört auch ASM eindeutig in den Bereich der Grid Infrastructure.

Cluster Registry und Voting Disk im ASM

Bisher hat man die Cluster Registry und die Voting Disk(s) jeweils auf ein separates Device, Partition oder ocfs gelegt. Zuerst sieht also die Integration dieser Teile in das ASM wie eine Vereinfachung aus. Allerdings birgt es auch ein paar Stolperfallen:

- Bei der ASM Konfiguration als „normal Redundancy“ werden jetzt statt 2 plötzlich 3 Kopien der OCR verlangt. Also auf jedes gespiegelte Storagedevice eine Kopie der OCR, aber wohin mit der dritten Kopie? Hierzu muss man einen zusätzlichen, reduzierten Spiegel des ASM Volumes anlegen, auf welchem die dritte Kopie gespeichert wird. Eine Vereinfachung ist hier nicht zu erkennen.
- Das ASM lässt sich nicht mehr stoppen, ohne den kompletten Clusterstack auf dem jeweiligen Knoten vollständig zu stoppen. Das ist logisch, wenn die OCR und die Voting Disk in einem ASM Volume liegen, verkompliziert aber die Arbeiten am ASM (Parameteränderungen, dismounten der Volumes oder ähnliches). Das erweist sich zumindest im Testbetrieb als sehr lästig.
- Bei Problemen mit der OCR oder der Voting Disk konnte man bisher vor irgendwelchen Reparaturexperimenten einfach einen DiskDump dieser Bereiche vornehmen (ohne dass der Cluster Stack laufen musste). Dies geht jetzt nicht mehr. Der Cluster Stack und das ASM muss laufen, um OCS oder Voting Disk sichern zu können. Was aber tun, wenn sich diese Aufgrund eines Fehlers nicht mehr starten lassen?

Block und Raw Devices für die OCR und Voting Disk werden nicht mehr unterstützt (außer beim Upgrade), jedoch sind Cluster Filesysteme oder NFS möglich.

ASM Cluster File System

Ein wirklich interessantes Feature. Ein nach außen nutzbares Filesystem, welches im ASM liegt, clusterfähig ist und alle Möglichkeiten des ASM als Volume Manager nutzen kann. Leider fehlen uns bisher noch die Ideen, wozu man es – außer als Oracle Home – noch sinnvoll nutzen kann. Wobei gegen die Nutzung als zentrales Oracle Home die Rolling Upgrade Fähigkeit des Clusters spricht.

Redundant Interconnect

Ab Oracle 11.2.0.2 können redundante Interconnect-Netzwerkverbindungen direkt von der Grid Infrastructure verwendet werden. Eine Vorab-Konfiguration der redundanten Interconnect-Interfaces (Bonding, Trunking, Teaming) mit allen damit verbundenen Problemen ist nicht mehr notwendig!

Bis zu vier Interconnects können damit automatisch für Load Balancing und High Availability konfiguriert werden.

SCAN als Erweiterung zur VIP

Der SCAN (Single Client Access Name) ist ein einheitlicher Hostname, mit dem alle Clients auf den Cluster zugreifen. Dieser Hostname wird im DNS (oder GNS) für mindestens eine und maximal drei IP-Adressen registriert. Auf den Clusterknoten wird ein separater SCAN-Listener gestartet, der die Clients beim Verbindungsaufbau an die passende VIP-Adresse weiterleitet.

Konkretes Beispiel:
Es existieren fünf Clusterknoten. Im DNS sind 3 IP-Adressen (aus dem gleichen Netzwerk wie die Serveradressen und die VIP-Adressen) registriert. Auf drei Clusterknoten läuft jeweils ein SCAN Listener, der auf eine dieser IP-Adressen lauscht. Welche Clusterknoten das konkret sind spielt keine Rolle.
Der Client möchte eine Verbindung zu einer Datenbank auf der Grid Infrastructure aufbauen. Dazu fragt er beim DNS den SCAN-Namen an. Vom DNS bekommt er (im RoundRobin Verfahren) zufällig eine der drei IP-Adressen mitgeteilt. Der Client baut eine Verbindung zu dieser IP-Adresse auf. Der dahinter lauschende SCAN-Listener teilt dem Client dann die (V)IP-Adresse mit, auf die er sich konkret verbinden soll.
Bisher mussten auf jedem Client alle verfügbaren VIP-Adressen der Datenbank konfiguriert werden. Wenn einer dieser Server aus dem Cluster entfernt wurde, existierte diese VIP-Adresse nicht mehr und alle Clients mussten angepasst werden. Dies erledigt sich jetzt durch den SCAN-Namen. Der SCAN-Listener weiß, welche Instanzen der Datenbank existieren und unter welchen VIP-Adressen sie zu erreichen sind.

Für dieses sehr sinnvolle Feature wird ein Client ab Version 11.2 benötigt, ältere Clients können dieses Verfahren nicht nutzen und müssen wie bisher auf die VIPs konfiguriert werden.

Cluster Time Synchronization Service

Die Zeit zwischen den beteiligten Clusterknoten muss immer synchron sein. Sollte bei der Installation keine NTP Zeitquelle gefunden werden, so wird automatisch CTSS konfiguriert, damit die Knoten mit einer einheitlichen Zeit arbeiten – egal wie weit diese von der realen Zeit abweicht.
Dieses Feature vermeidet sicherlich Fehler bei der Installation, allerdings kann es keine echte NTP Zeitquelle ersetzen. Hier sollte lieber vorab der NTP-Server korrekt konfiguriert werden.

Grid Naming Service

Um den Grid Naming Service (GNS) nutzen zu können, muss dem Cluster vom DNS eine Subdomain inklusive einem Adressbereich zur eigenen Verwaltung zugewiesen werden. Der Grid Naming Service verwaltet also seine eigene Subdomain (z.B. grid.mydomain.com).
Innerhalb dieser Subdomain weist der GNS über DHCP allen beteiligten Knoten ihre IP-Adressen zu, auch inklusive der Cluster Interconnect Adressen. Das bedeutet, dass man sich selbst nicht mehr um die IP-Adressen kümmern muss.

Dies bedeutet jedoch auch, dass man nicht mehr weiß, welcher Knoten gerade welche IP-Adresse hat. Dadurch wird die Fehlersuche in der Praxis nicht wirklich einfacher. Nur mit Hilfe des Grid Control als zentraler Informationsquelle käme man hier weiter. Leider weist Grid Control gerade in diesem Bereich (wo läuft im Moment welcher SCAN Listener, welcher Server hat welche IP-Adresse) noch viel Raum für Verbesserungen auf. Wir haben daher dieses Feature nach ersten Experimenten nicht genutzt, zumindest so lange, bis die vollständige und fehlerfreie Unterstützung im Grid Control verfügbar ist.

Policy Based Cluster

Im Gegensatz zum Administrator verwalteten Cluster werden die Server im Policy verwalteten Cluster in Server Pools unterteilt. Eine Datenbank (genauer ein Service) wird genau einem Server Pool zugeordnet.
Jeder Server kann zwar zu mehreren Server Pools gehören, aber nur zu genau einem zu einer Zeit.

Für jeden Server Pool wird festgelegt, welche Mindest- und Maximalzahl an Servern in diesem Pool benötigt werden und welche Priorität der Pool hat. Anhand dieser Policy errechnet die Clusterware, wie viel Server aktuell in einem Pool vorhanden sind.
Ein einfaches Beispiel:
Es gibt 5 Server im Cluster. Dazu wurden drei Server Pools definiert. Pool A hat eine Mindestgröße von zwei und eine Maximalgröße von vier Servern sowie eine Priorität von 50 (je größer die Zahl desto höher die Priorität). Pool B hat eine Minimalgröße von 2 Servern, eine Maximalgröße von 3 Servern und eine Priorität von 30. Pool C hat eine Minimalgröße von einem Server, eine Maximalgröße von 3 Servern und eine Priorität von 10.
Dadurch werden die Server wie folgt eingeteilt:
Pool A besteht aus 2 Servern (=Minimum).
Pool B besteht ebenfalls aus 2 Servern (=Minimum).
Pool C besteht aus einem Server (=Minimum).
Damit sind alle 5 Server in Pools eingeteilt. Die Instanzen auf den Servern werden entsprechend Ihrer Poolzugehörigkeit gestartet.
Sollte jetzt ein Server ausfallen geschieht etwas Unerwartetes: Da Pool A und Pool B eine höhere Priorität haben als Pool C und für beide Pools eine Mindestgröße von jeweils 2 Servern definiert ist, bleibt für Pool C kein Server mehr übrig, da dieser eine niedrigere Priorität hat als die anderen zwei Pools. Auch das definierte Minimum von einem Server zieht jetzt nicht mehr – der Server Pool C wird einfach heruntergefahren, da kein Server mehr verfügbar ist. Damit sind auch alle Instanzen, die diesem Server Pool zugeordnet waren, plötzlich nicht mehr verfügbar.

Dies ist nicht das Verhalten, welches sich ein Administrator vorstellt. Ein Server im Cluster fällt aus – eigentlich ein vollkommen unspektakuläres Ereignis. dafür ist ein Cluster schließlich gedacht. Aber plötzlich sind einige Services für die User nicht mehr erreichbar, da diese einfach heruntergefahren wurden. Und aus einem harmlosen Ereignis entsteht plötzlich Hektik.
Ich habe hier noch ein relativ einfaches Beispiel gewählt – in komplexeren Umgebungen kann dies relativ schnell sehr unübersichtlich werden. Und man kann nur noch sehr schlecht überblicken, welcher Ausfall welche Auswirkungen hätte. Das empfinde ich als ziemlich kontraproduktiv für eine Hochverfügbarkeitslösung – hierbei sollte man immer souverän den Überblick behalten können.

Insofern werden wir dieses Feature – obwohl wir es sehr gerne genutzt hätten – erst einmal nicht verwenden, da wir die Auswirkungen im Einzelfall nicht überblicken können. Stattdessen arbeiten wir weiter mit einem Administrator verwalteten Cluster und teilen die Ressourcen selbst zu.

Beurteilung der Neuerungen

Es gibt einige Neuerungen, die eine logische und konsequente Weiterentwicklung darstellen und durchaus positiv zu bewerten sind. Insbesondere die Integration von ASM in die Grid Infrastructure, der redundante Interconnect und der SCAN sind eine prima Sache.

Etwas kritischer sehe ich noch den GNS – hier fehlt die Unterstützung des Grid Control. In Zukunft dürfte es allerdings durchaus Standard sein – ich empfinde es als ein ähnliches Unbehagen wie man es seinerzeit bei OMF hatte – inzwischen nutzen wir fast ausschließlich OMF.
Noch unschlüssig bin ich beim Policy Based Cluster. Meine Vorstellung einer dynamischen Ressourcenverwaltung sieht schlichtweg anders aus als diese Implementierung.

Kontaktadresse:

Jochen Kutscheruk
merlin.zwo InfoDesign GmbH & Co. KG
Büro Karlsruhe
Taglöhnergärten 43
D-76228 Karlsruhe

Telefon: +49 (0) 721-79 071 71
Fax: +49 (0) 721-79 071 98
E-Mail jochen.kutscheruk@merlin-zwo.de
Internet: www.merlin-zwo.de