Deseaster Recovery bei Grid Infrastructure 11.2 mit 2 Rechenzentren

Thorsten Bruhns
Opitz-Consulting Bad Homburg GmbH

Schlüsselworte: Grid Infrastructure, CRS, RAC, High Availibility

Unternehmen bauen zunehmend häufiger Clustersysteme auf, um die Verfügbarkeit von Applikationen zu verbessern. Dabei erfolgt die Installation derartiger Systeme häufig in 2 getrennten Rechenzentren bzw. Brandabschnitten mit je einem Storagesystem pro Rechenzentrum.
Für einen sicheren Betrieb sind einige Punkte zu beachten, damit beim Ausfall eines Rechenzentrums der Betrieb möglichst schnell wieder aufgenommen werden kann.

Historie
Oracle hat mit dem Release 11.2 die Clusterware (CRS) in Grid Infrastructure (GI) umbenannt. Diese Änderung betrifft nicht nur den Namen sondern auch den Clusterstack, da es wesentliche Strukturänderungen gab.
Folgenden Punkte sind hinsichtlich eines Deseaster-Recoveries von Bedeutung:

- ASM ist Bestandteil der GI geworden.
- Voting/OCR und SPFile vom ASM liegen in einer Diskgruppe
  => keine getrennten Block-Devices für OCR/Voting
  => Henne/Ei-Problem, wie kann Voting vom cssd ohne ASM gelesen werden?
  Wie kann ASM ohne cssd gestartet werden?

Annahmen
Es wird davon ausgegangen, das nur 2 Rechenzentren mit jeweils 1 Storagesystem zur Verfügung stehen. Ein dedizierter NFS-Server für das 3 Voting existiert **NICHT**. Die Spiegelung der Daten erfolgt mittels ASM. (keine SAN-Virtualisierung)
Es sind 3 Votings vorhanden.
OCR kann mit in die Votingdiskgruppe gelegt werden. Optional kann auch eine zusätzliche Diskgruppe als Mirror für OCR konfiguriert werden.

Probleme im Deseasterfall
Aus den Annahmen ergeben sich Probleme beim Ausfall eines Rechenzentrums. Die GI benötigt immer eine Mehrheit an lesbaren Votings, um starten zu können. Dieser Zustand ist jedoch nicht immer gegeben, weil 3 Votings über 2 Standorten verteil werden müssen. Eein ständiger Zugriff auf alle Votings stellt sicher, das alle beteiligten Knoten sich sehen können– parallel zur Netzwerkkommunikation über den Interconnect.
Sind im Deseasterfall weniger als 50% der Votings im noch aktiven Rechentrum verfügbar, dann wird automatisch ein Reset des verbliebenen Clusterknoten erfolgen.
Beim Neustart kann die GI nicht starten, da die Mehrheit der Votings nicht erreichbar ist.

Diese Situation erfordert einen manuellen Eingriff in die Konfiguration, um wieder ein lauffähiges System zu erhalten.

In der Praxis sind derartige Systeme häufiger anzutreffen als man erwarten würde, da nur selten ein 3. Rechenzentrum für die Installation eines NFS-Server zur Auslagerung eines Votings vorhanden ist.

Vorbereitende Maßnahmen
Vor der Installation werden pro Rechenzentrum 2 Votingdevices im Storage konfiguriert. Hier sollte darauf geachtet werden, das die Namen der Devices eindeutig zum Rechenzentrum zugeordnet werden können, damit klar sichtbar ist, wer beim Split-Brain ‚überlebt'.
Die Installation erfolgt so, das 2 Votings im nicht bevorzugten Rechenzentrum liegen, da nach Abschluß der Installation eine Konfigurationsänderung der Votings erfolgt, die einen Wechsel des bevorzugten Rechenzentrums zur Folge hat.
Das 4.– derzeitig nicht verwendete – Voting wird mit der Option ‚alter diskgroup <diskgroup> add **QUORUM** …' hinzugefügt und das 2. Voting aus dem anderen Rechenzentrum entfernt. So ist sicher gestellt, das die Diskgruppe von Voting 3 Failuregruppen hat, von denen eine ausschließlich für Voting verwendet wird. Sollte die Diskgruppe die mit ‚QUORUM' markierte und eine der beiden verbliebenen Failuregruppe verlieren, dann ist die Diskgruppe selbst noch nutzbar, weil ASM die Spiegelung durch die ‚NORMAL REDUNDANCY' einhalten kann. Diese Bedingung ist sehr wichtig, weil im Rahmen der Installation die OCR und das SPFile vom ASM in der Diskgruppe abgelegt werden. Darüber hinaus wäre die Diskgruppe beim Ausfall des bevorzugten Rechenzentrums zerstört und die notwendige Rekonfigurationstätigkeiten erheblich aufwendiger.

Deseasterfall: Ausfall des nicht bevorzugten Rechenzentrums
Hier gibt es nichts zu beachten, da lediglich 1 Failuregruppe von Voting und der Spiegel der ASM-Redundancy von der OCR und dem SPFile verloren geht.

Der Cluster wird ohne Eingriff im 2. Rechenzentrum weiter laufen!

Deseasterfall: Ausfall des bevorzugten Rechenzentrums
Hier wird der verbliebene Clusterknoten einen Reset durchführen, weil er über den Interconnect keinen Clusterpartner mehr sieht und keine Mehrheit bei den Votings erreichen kann. Diese fehlende Mehrheit verhindert den anschließenden Neustart der GI.
Die GI muß nun manuell im ‚exclusive' Modus gestartet werden, wo der cssd-Prozeß prüft, ob er auf dem Interconnect einen aktiven Clusterknoten sieht. Im Anschluß kann die ASM-Resource gestartet werden, so das eine Reparatur der ASM-Diskgruppe möglich wird, damit der cssd wieder mehrheitsfähige Votings sieht. Für diese Reparatur wird das im SAN noch freie LUN der Diskgruppe als weitere ‚QUORUM'-Failuregruppe hinzugefügt. Hier ist wieder darauf zu achten, das die ‚QUORUM'-Option verwendet wird, damit nach der Reparatur des 2. Rechenzentrums der notwendige Spiegel für OCR und SPFile auch sicher über die Rechenzentren verteilt wird.

Alle notwendigen Details sowie entsprechende Bilder zur Illustration werden im Vortrag erläutert.

Kontaktadresse:

Thorsten Bruhns
Opitz-Consulting Bad Homburg GmbH
Kaiser-Friedrich-Promenade 93-95
D-61348 Bad Homburg

Telefon: +49 (0) 6172 – 66 26 0 1541
Fax: +49 (0) 6172 – 66 26 0 4541

E-Mail thorsten.bruhns@opitz-consulting.de
Internet: www.opitz-consulting.com