Aufbau eines Exadata-Ecosystems bei der Otto Group

Björn Gauerke
Otto GmbH & Co KG
Hamburg

Schlüsselworte
Exadata, Migration, ZFS Backup Appliance, HA, RAC-DG-RAC, MAA

Einleitung
Dieser Vortrag gibt einen Einblick in die technischen und organisatorischen Abläufe bei dem Aufbau eines Exadata "Ecosystems" für das ERP-System der Otto Group.

Nach einem erfolgreichen Proof-of-Concept für das ERP-System der Otto Group wurde die Entscheidung getroffen, die bestehende Infrastruktur auf HP-Superdome durch eine hochverfügbare Exadata-Umgebung zu ersetzen.

Die Migration wurde in der vorgegebenen Zeit durchgeführt und diverse neue Konzepte im Umfeld Backup/Recovery und Hochverfügbarkeit wurden fristgemäß umgesetzt.

Proof of Concept
Für den Proof of Concept stellte uns Oracle ein Halfrack (X2-2) zur Verfügung, auf dem wir die Software unserer ERP-Applikation installierten und migrierten.

Wir führten einen kompletten Integrationstest mit unserer Test-Umgebung durch und ermittelten die Laufzeiten der Batch-Komponenten. Außerdem wurden die Antwortzeiten eines Online-Lasttests unter Extrembedingungen gemessen.

Wir erzielten folgende Ergebnisse:

- Die Laufzeiten der Batchprogramme waren schneller, als mit der aktuellen Hardware
- Im kritischen Pfad gab es einige Verarbeitungs-Pakete mit erheblich kürzerer Laufzeit, andere zeigten keinen Performancegewinn

In einem weiteren Testzyklus optimierte die Otto-Software-Entwicklung gemeinsam mit Oracle einzelne Teilkomponenten des kritischen Pfades, um das Potential von Software-Anpassungen speziell für Exadata aufzuzeigen:

- Frequentierte Tabellen wurden im Storage-Server Flash fixiert
- Durch Entfernung von Indizes wurden Storage-Scans forciert

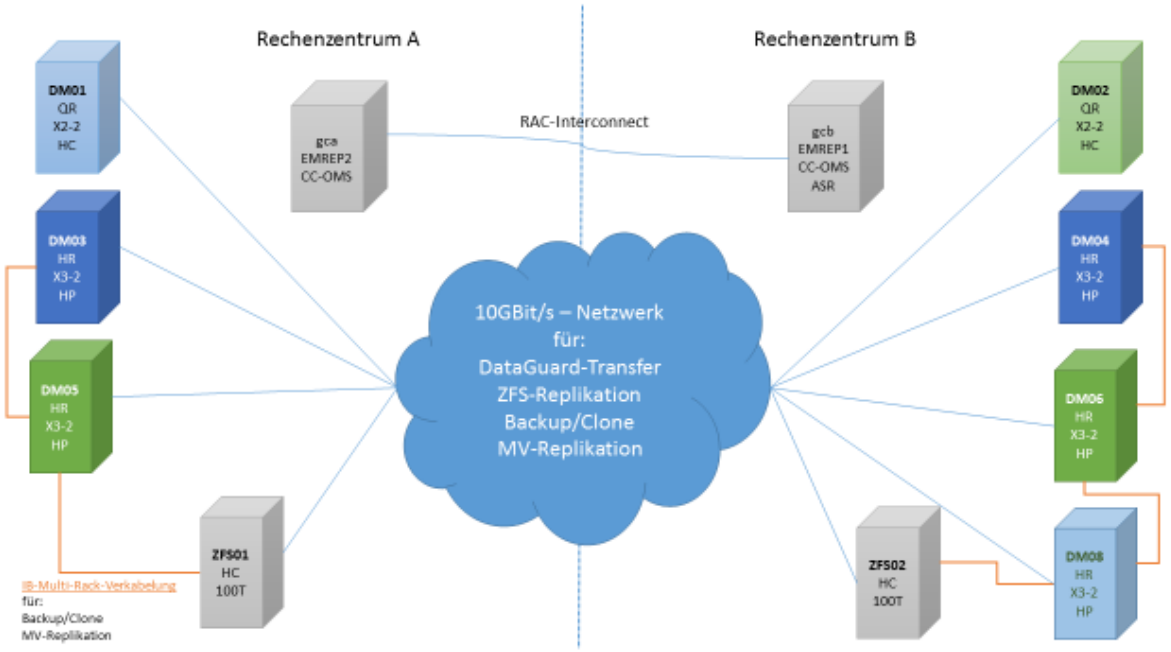
Das Optimierungspotential war sehr hoch. Zusammen mit der Otto-Software-Entwicklung wurde beurteilt, ob und in welchem Zeitraum die Optimierungen in der Releaseplanung berücksichtigt werden können.

Eine mittelfristige Umsetzung von Exadata-Optimierungen im normalen Software-Entwicklungs-Zyklus wurde als machbar beurteilt.

Abzulösende Infrastruktur
Die zur Ablösung anstehende Infrastruktur bestand aus einem RAC-Cluster auf HP-Superdome-Maschinen mit angeschlossenem EMC Storage. Der EMC-Storage wurde über Host-Base-Mirror über die RZs gespiegelt.

Diverse produktionsgleiche Testsysteme wurden über BCV-Split-Technologie aufgebaut.

Exadata-Lösung (RAC-DG-RAC)
Die hochverfügbare Exadata-Lösung besteht im Kern aus einer RAC-DataGuard-RAC Umgebung. Backup/Recovery und Cloning wird durch eine hochverfügbare ZFS-Backup-Appliance-Lösung gewährleistet. Ein dediziertes Netzwerk ermöglicht die High-Speed-Kommunikation der Systeme rechenzentrumsübergreifend. Eine ausfallsichere CloudControl-Installation sorgt für die Überwachung und das administrative Jobnetz.



Migration
Für die Migration von HP zu Exadata wählten wir Datapump-Ex- und -Import. Aus dem Business hatten wir die Vorgabe, die Datenbank innerhalb einer maximal 6 Stunden dauernden Downtime auf Exadata zur Verfügung zu stellen. Dabei mussten wir den folgenden Herausforderungen begegnen:

- Begrenzter Speicherplatz auf dem EMC-Storage für den Export
- Beschränkte Netzwerkverbindung zwischen Superdome und Exadata auf 1 GBit/s
- Export/Import ohne Anpassungen und Optimierungen würde das Zeitfenster um Faktoren sprengen - dies war bereits aus diversen PoCs bekannt

Unsere Lösung:

- Hochparallelisierter, komprimierter Export in die FRA (ASM) auf dem HP-System
- Transfer der Datapump-Dumps (ASM-ASM) unter Ausnutzung der maximalen Netzwerk-Bandbreite

- Import in mehreren Teilschritten, um eine maximale Parallelisierung zu gewährleisten und die neue Infrastruktur maximal zu nutzen

Die ersten beiden Schritte waren schnell und maximal zu optimieren.

- Export mit einem Parallelitätsgrad von 32
- Größe des Dumps von ca. 220 GB
- Dauer des Exports ca. 120 Minuten

Unter Berücksichtigung der Dump-Größe von 220 GB und der vorhandenen Bandbreite von 1 GBit/s ergab sich eine optimale Transferzeit von knapp über einer halben Stunde. Dies erreichten wir in den ersten Tests und weiteren Tests sehr stabil.

Nachdem für die ersten beiden Schritte 2,5 Stunden benötigt wurden, blieben noch 3,5 Stunden für den Import der im unkomprimierten Zustand ca. 3,5 TB umfassenden Datenbank. Bei der Datenbank handelt es sich um ein komplexes, hochnormalisiertes Datenmodell mit teilweise umfangreicher Ausnutzung von Oracle-Features.

Den kompletten Export/Transfer/Import Vorgang führten wir vor der eigentlichen Migration drei Mal unter Livebedingungen durch, um die Testumgebungen für die Software-Entwicklung zur Verfügung zu stellen. Darüber hinaus haben wir - besonders den Import - weitere Male durchgeführt, um den Zeitaufwand und die Stabilität zu optimieren.

Erfolgsfaktoren für den Import in die Exadata:

- Alleiniger Import der Tabellen (Parallelitätsgrad=100)
- Erstellen eines Indexfiles
- Anpassen des Indexfiles, um die Indizes "unusable" anzulegen
- Anlegen der unusable Indizes
- Hochparallelisiertes Rebuild aller unusable Indizes mittels eines selbstentwickelten Maintenance-Frameworks
- Erstellen der Constraints (no-validate)

Aufgrund dieser Optimierungen waren wir in der Lage, die Datenbank innerhalb der vorgegebenen Zeit zu übergeben.

Aufbau von Testumgebungen mittels einer hochverfügbaren ZFS-Backup-Appliance
Sowohl das Host-Ablösungs-System, als auch das ERP-System benötigen auf den Exadata diverse Test-Systeme mit Produktions-Datenvolumen.

Da uns keine EMC-BCV-Mirror-Technologie zur Verfügung steht, wurde dieser Prozess für Exadata neu definiert. Um einen hochperformanten Clone-Prozeß zu etablieren, wählten wir RMAN-Technologie und als Storage-Einheit eine hochverfügbare ZFS-Backup-Appliance.

Wir verwenden das Backup der Produktion, um aus diesem die jeweiligen Test-Systeme zum gewünschten Zeitpunkt wiederherzustellen.

Ein entsprechendes Framework realisiert auch hier, dass Jobs lediglich über CloudControl angesteuert werden.

Appliance-Ansatz
Drittsoftware auf den Exadata-Computenodes
Um für die Exadatasysteme eine möglichst geringe Abhängigkeit von Drittsoftware zu gewährleisten, haben wir uns entschieden, diese nach Möglichkeit gar nicht zuzulassen.

Hierbei lösten wir Abhängigkeiten folgendermaßen auf:

Datenbank-Backup
Da für den Aufbau der Test-Systeme die ZFS-Appliance verwendet werden sollte, war auf den Compute-Nodes kein separater Backup-Agent notwendig. Das Datenbank-Backup wird über CloudControl-Jobs direkt auf die ZFS geschrieben und dort für die Retentionzeit vorgehalten.

Eine ggf. benötigte Langfristsicherung kann direkt von der ZFS-Appliance über Backup-Server auf Tape erfolgen.

Exadata-Monitoring
Da für das Exadata-Monitoring CloudControl gesetzt ist, wollten wir vermeiden, weitere Monitoring-Agent-Software auf den Computenodes zu installieren. Dies lösten wir durch eine generische Schnittstelle zwischen CloudControl-Server, Haus-Monitoring und dem angeschlossenen RZ-Leitstand. Somit konnte die Installation weiterer Monitoring-Software vermieden und die Verantwortung für das Monitoring auf CloudControl konzentriert werden.

Job- und Batchverarbeitung
Auf der alten Infrastruktur liefen diverse administrative aber auch fachliche Scripts über Cronjobs oder direkt auf einem Knoten gesteuert durch das Batchsystem.

Hier gab es früher bereits das Problem, dass diese Jobs teilweise nicht Cluster-Aware waren.

Die Jobs trennten wir auf:

- Fachliche Scripts wurden auf bereits vorhandene Batch-Process-Blades ausgelagert
- Administrative Jobs wurden auf CloudControl migriert

Als Ergebnis gibt es auf den Computenodes keine lokal ausgeführten Scripts mehr. Sämtliche alten und neuen Admin-Scripts werden über CloudControl gesteuert und überwacht.

Computenode Backup und Recovery
Das Backup der Exadata-Compute-Nodes führen wir ebenfalls auf den ZFS-Systemen durch. Basis des Compute-Backups ist das Dokument: B&R Konzept nach Exadata „Owner‘s Guide“ (E13874-28) – Kapitel „Recovering a Linux-Based Database Server Using the Most-Recent Backup”.

Für das Backup müssen grob die folgenden Schritte durchgeführt werden:

- Mounten eines Shares der ZFS an den Computenode
- Erstellen von LVM-Snapshots der Filesysteme / und /u01
- Sichern der Filesystem-Snapshots und des /boot Filesystems mittels tar auf die ZFS

Für den Restore müssen die folgenden Schritte durchgeführt werden:

- Starten des Computenodes via diag.iso - diese Datei findet sich auf jedem Computenode unter /opt/oracle.cellos/diag.iso und ist somit auch im Tarfile des Backups vorhanden
- Nach dem Booten wählt man den Restore-Modus aus

- Nach der Eingabe des Rescue-Passwortes gibt man den kompletten NFS-Pfad zum TAR-File an
- Als letztes wählt man ein Interface zwischen eth0 und eth3 aus, über das die Verbindung zum NFS-Share für den Restorevorgang aufgebaut werden soll

Beim Restore-Test haben wir die folgenden Erfahrungen gemacht:

- Es ist gar nicht so einfach, von Oracle entsprechende Systemplatten zu erstehen
- Unsere ZFS ist nur über die Interfaces eth4 und eth5 zu erreichen, dies bietet einem der Restore-Prozess aber nicht an
- Wir konnten also problemlos ein Backup durchführen, kämen im K-Fall aber nicht an dieses heran, deswegen führten wir zwei Restore-Versuche durch:
    - Kopieren des TAR-Files auf einen anderen Computenode und erstellen eines NFS-Exportes. Somit waren wir in der Lage über das Management-LAN auf die Backup-Datei zuzugreifen und erfolgreich wiederherzustellen.
    - Nach einem zwischenzeitlichen Update der ZFS-Appliance waren wir in der Lage über ein zusätzliches virtuelles Interface die ZFS über das gemeinsame Management-LAN zu erreichen. Auch hier war ein Restore erfolgreich.

Abgesehen von den oben aufgeführten Einschränkungen ist der in der Dokumentation beschriebene Ablauf erfolgreich durchführbar. Die Dauer des eigentlichen Restore-Vorgangs hängt von Anzahl und Größe der Dateien ab, die auf dem jeweiligen Computenode vorhanden sind. Auf einem frisch installierten Computenode (mit sehr geringem "diag"-Datenvolumen) ist dieser Vorgang in weit weniger als einer Stunde abgeschlossen.

Organisatorische Aufstellung

Um eine einheitliche Betreuung beginnend bei der Hardware bis zum Datenbank-Stack zu gewährleisten, bildeten wir ein "Appliance Team". So bündeln wir die Kompetenzen und bieten bis zur Ebene der Datenbank einen Ansprechpartner.

Die Grundzusammenstellung des Teams bestand aus einem System-Administrator und einem DBA. Mit der zunehmenden Anzahl der zu betreuenden Systeme (aktuell 5xHR, 2xQR, 2xZFS, 2xCloudControl) erweiterten wir das Team um einen Organisator und zwei weitere "Appliance-Administratoren". Durch die gemachten Erfahrungen und entsprechender Dokumentation ist dieses Team nun in der Lage, innerhalb kurzer Zeit auf Anforderungen aus den Fachbereichen zu reagieren. Das "Onboarding" neuer Exadata-Racks ist - besonders auch aufgrund der etablierten Standards - innerhalb kürzester Zeit möglich.

Die folgende Timeline zeigt diverse geplante und ungeplante Tasks - über die Gewährleistung des DB-Betriebes hinaus - die bisher von uns bewältigt werden mussten.

Timeline

- 1. HJ 2012 - Aufbau von 2 Quarterracks für das Host-Ablösungs-Projekt
- 1. HJ 2012 - Aufbau der Monitoring-Infrastruktur CloudControl 12c
- 2. HJ 2012 - Proof of Concept mit dem Otto Group-ERP-System
- Januar 2013 - Aufbau von zwei Halfracks für das Otto Group-ERP-System
- Januar 2013 - Aufbau von zwei weiteren Quarterracks für das Host-Ablösungs-Projekt
- März/April 2013 - Aufbau von zwei ZFS-Backup-Appliances
- September 2013 - Migration des ERP-Systems
- Dezember 2013 - Aufbau eines weiteren Halfracks für das Host-Ablösungs-Projekt
- Januar 2014 - Erweiterung von zwei QR des Host-Ablösungs-Projektes auf Halfrack

- März 2014 - Umbau von zwei QR von High-Performance auf High-Capacity

Weitere Aktionen:

- Diverse DB Umbauten/Umzüge unter Produktionsbedingungen
- Online-Erweiterung der ZFS von 50TB auf 100TB Netto
- Netzwerkumbau im laufenden Betrieb auf eine eigene 10 GBit/s Umgebung für Backup/DG/Replikation
- Netzwerkumbau im laufenden Betrieb auf ein separates Management-LAN
- Koordinierung bzw. Durchführung normaler Hardware-Aktionen: DIMM-Speicher/Platten/PDU/FlashCard-Tausch
- Diverse CloudControl-Upgrades (12.1.0.2/3/4)

Kontaktadresse:

Björn Gauerke
OTTO (GmbH & Co KG)
Werner-Otto-Straße 1-7
D-22179 Hamburg

Telefon:	+49 (0) 40-6461 4156
Fax:	+49 (0) 40-6464 4156
E-Mail	bjoern.gauerke@ottogroup.com
Internet:	www.ottogroup.com