Konsolidierungsplattform Exadata – PoC-Erfahrungsbericht zeigt Licht und Schatten

Gregor Büchner
T-Systems International GmbH
Münster

Uwe Simon
T-Systems International GmbH
Bonn

Joachim Dietsch
Oracle Deutschland B.V. & Co.KG
Dreieich

Schlüsselworte
Datenbankserver-Konsolidierung, Exadata, OVM, PoC, Erfahrungsbericht

Einleitung

Die Telekom-IT ist der interne Dienstleister der Deutschen Telekom. Sie bündelt seit 2012 alle IT-Einheiten des Konzerns und ist verantwortlich für das komplette IT-Service-Portfolio. Die Telekom-IT nutzt die Infrastruktur- und Cloud-Services der T-Systems Market Unit (im Wesentlichen das externe Geschäft mit Großkunden).

Die Telekom-IT betreibt mehr als 1000 Oracle-Datenbanken mit einem Volumen von mehreren Peta-Byte verteilt über mehrere Rechenzentren auf unterschiedlicher Hardware (Power/Sparc/x86) mit verschiedenen Betriebssystemen (AIX, Solaris, Linux). Für eine anstehenden Datacenter-Konsolidierung wird eine kostengünstige, leistungsfähige Private-Cloud-Plattform für Oracle Datenbanken unter Berücksichtigung unserer Security-Anforderungen gesucht. Nach einer Oracle Insight Studie hat sich die Telekom IT für ein PoC mit einer Exadata X5-2 entschieden. OracleVirtualMachine (OVM) wurde eingeplant, um die Datenbanken zu isolieren und um dem core-basierten Lizenzmodell gerecht zu werden. IO-Ressourcen-Management schützt im Fall von IO-Engpässen produktive von non-produktiven Datenbanken. Die Elastic-Configuration ermöglicht ein demand-basiertes Wachstum der Exadata.

Die Idee

Wie in der Einleitung beschrieben, hat die Telekom-IT eine Datacenter-Konsolidierung beschlossen. Am neuen Standort wurden dafür neue Plattformen etabliert – mit dem Ziel, für Applikationen der Telekom-IT eine standardisierte, Security konforme Umgebung anbieten zu können.

In den heutigen Datacentern wurden ein Großteil der Server noch individuell konfiguriert und betrieben. Server von unterschiedlichen Herstellern kamen zum Einsatz. Die Installationen waren entsprechend pro Plattform standardisiert und es gab kein einheitliches Betriebsmodell. Das erschwerte Automatisierung und verlangsamte Bereitstellungszeiten entsprechend.

Abhilfe sollen neu aufgebaute Konsolidierungs-Plattformen für unterschiedliche Anforderungen mit unterschiedlichen Betriebssystemen/Prozessortypen (AIX auf Power/Solaris auf SPARC/Linux/Windows auf x86) schaffen. Alle diese Plattformen unterstützen Virtualisierungs-Technologien. Auf der AIX- und der Solaris-Plattform sind auch Oracle-Datenbank-Installationen umsetzbar. Die x86-Plattform lässt das im Moment noch nicht zu – aus zwei Gründen: zum einen, weil mit Oracle eine core-basierte Lizenzvereinbarung besteht und zum anderen, weil darauf VMware zur Virtualisierung läuft.

Die beschriebenen Plattformen dienen zur Konsolidierung von Web- und Applikations-Servern. Oracle-Datenbanken laufen zwar auf AIX und Solaris, allerdings sind insbesondere der angeschlossene SAN- und NAS-Storage sowie die Backup-Anbindung bei steigenden Datenvolumina nicht optimal. Das gleiche gilt für das Betriebsmodell dieser Plattformen (aus Sicht des DBA), da den OS- und Storage-Betrieb unterschiedliche Teams erbringen. Dies beeinträchtigt Bereitstellungszeiten und wirkt sich bei Incidents im laufenden Betrieb aus. Aus kommerzieller Sicht ist zu berücksichtigen, dass IBM Power Cores einen Oracle-Lizenz-Faktor von eins haben (gegenüber 0,5 bei Exadata), und dass Power- und SPARC-Server aus der High-End-Klasse eingesetzt werden.

So entstand der Gedanke eine optimierte Plattform für Oracle-Datenbanken zu entwickeln, die den Ansprüchen an Konsolidierung, Security und an ein standardisiertes vereinfachtes Betriebsmodell genügen. Unterm Strich sollte die neue Plattform dann natürlich Kosten sparen und die Bereitstellungszeiten deutlich verkürzen.

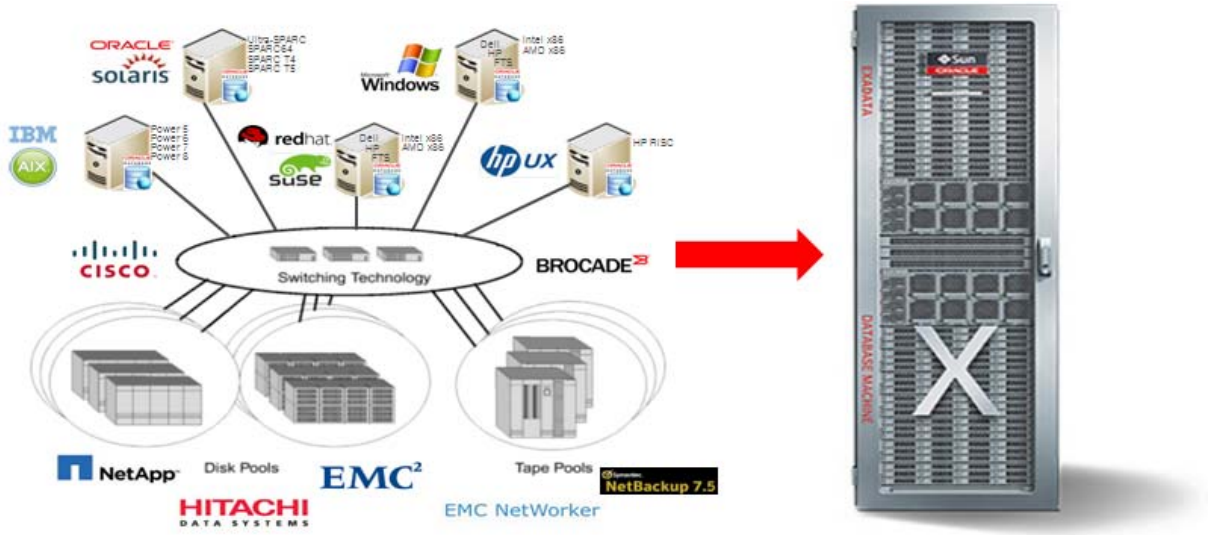

Abb. 1: Konsolidierung und Standardisierung mit Oracle Exadata

Der Business Case

Mit einer sich ständig ändernden IT gewinnt die Business-Case-Betrachtung immer mehr an Bedeutung, insbesondere in der frühen Projekt-Phase. Das Oracle-Insight-Programm unterstützt das, weil es die Business-Anforderungen analysiert und daraus eine optimale Architektur entwickelt.
Fokus einer Insight-Studie ist primär die Analyse des „*Why*“, also die Business-Begründung. Ein weiterer Bereich, das „*What*“ - also die Future-State-Architektur - wurde durch ein erstes High-Level Sizing ermittelt. Auch der dritte Bereich, das „*How*“, also die Roadmap, entstand als Skizze anhand von Best-Practice-Migrations-Szenarien.

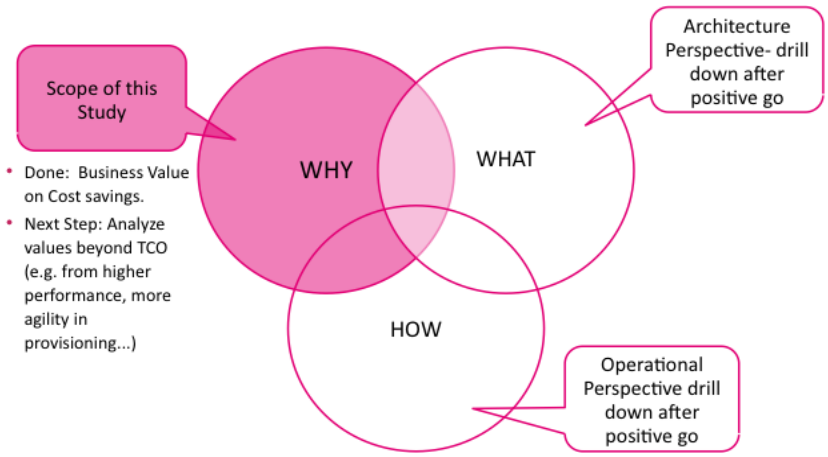

Abb. 2: Fokus der Insight-Studie

Das Ziel des Projekts war von Anfang an klar definiert: Höhere Agilität, mindestens gleichbleibende Performance und gleichzeitig geringere Kosten.

Der Current Mode analysierte zu Beginn eine potentielle installierte Basis von über 2.100 CPU-Cores, welche sich auf rund 200 physikalische Server mit folgenden Plattformen verteilten:
(Listung nach Häufigkeit)

1. AIX / Power
2. Solaris / SPARC
3. LINUX / x86
4. HP-UX / Itanium

Zur Aufschlüsselung der Baseline im Current Mode wurden die wichtigsten Kosten-Faktoren sowohl in absoluten Werten als auch in prozentualer Verteilung ermittelt:

- Managed Storage Costs
- Managed DB-Server Costs
- Infrastructure Costs
- DB SW-Maintenance Costs

Diese Kosten-Struktur spiegelt die Ausgangslage und gleichzeitig die „*do nothing*“ Strategie (ebenfalls eine Option) wieder.
Zum Vergleich wurden die genannten Kosten-Faktoren den bereits vorhandenen IaaS-Plattformen mit ihren unterschiedlichen Architekturen zugeordnet:

- IT-Serverfarm AIX
- IT-Serverfarm Solaris
- IT-Serverfarm x86

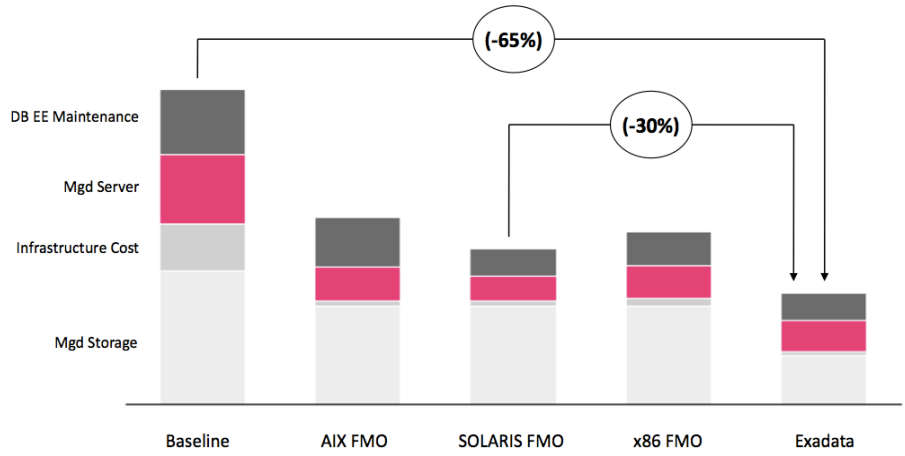

Abb. 3: Kostenverteilung Current State (Baseline) vs. mögliche Varianten für Future State

Bei einer geplanten Projekt-Laufzeit von 4 Jahren konnte die Exadata-Architektur bei allen vier Kosten-Bereichen deutliche, zusätzliche Einspar-Potenziale aufzeigen:

- 65% Gesamtersparnis gegenüber dem Current State und
- nahezu 30% gegenüber der günstigsten Implementierung auf Basis einer IT-Serverfarm

Da diese DBaaS-Architektur eine Universal-Plattform für alle anfallenden Workloads darstellen sollte (OLTP, DWH), wurde für das Sizing der Exadata eine generische Formel angewendet.
„SQL Offloading“ ermöglicht bei den Storage-Server eine Auslagerung von einigen Funktionen auf die CPUs der Storage Nodes, was insbesondere bei DWHs zu Beschleunigung bzw. Entlastung führt. Das Sizing hat dies berücksichtigt und bei dem DB-Servern die benötigten CPUs bzw. Cores um 30% reduziert.

Abgerundet durch eine Ramp-Up-Kurve auf Basis einer Exadata-Architektur ist ein ROI in bereits 9 Monaten darstellbar.

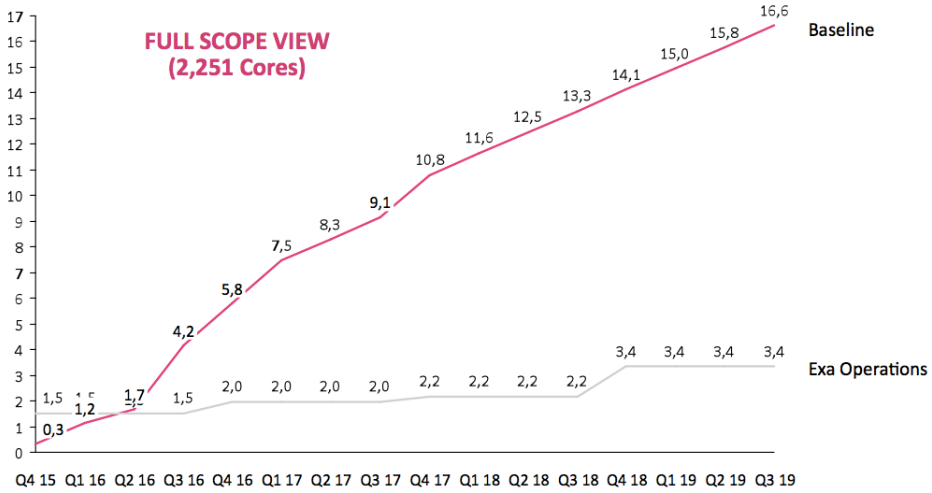

Abb. 4:Business Case Einsparungspotential

Die Planung
Nachdem die Anforderungen an die neue Oracle-DB-Plattform feststanden, wurde eine Architektur entworfen, die folgende Kernpunkte bei Bedarf erfüllen kann:

- Hochverfügbarkeit
- Desaster-Recovery-Fähigkeit
- Abbildung verschiedener Security-Zonen
- Isolation von Produktions- und Nicht-Produktions-Datenbanken

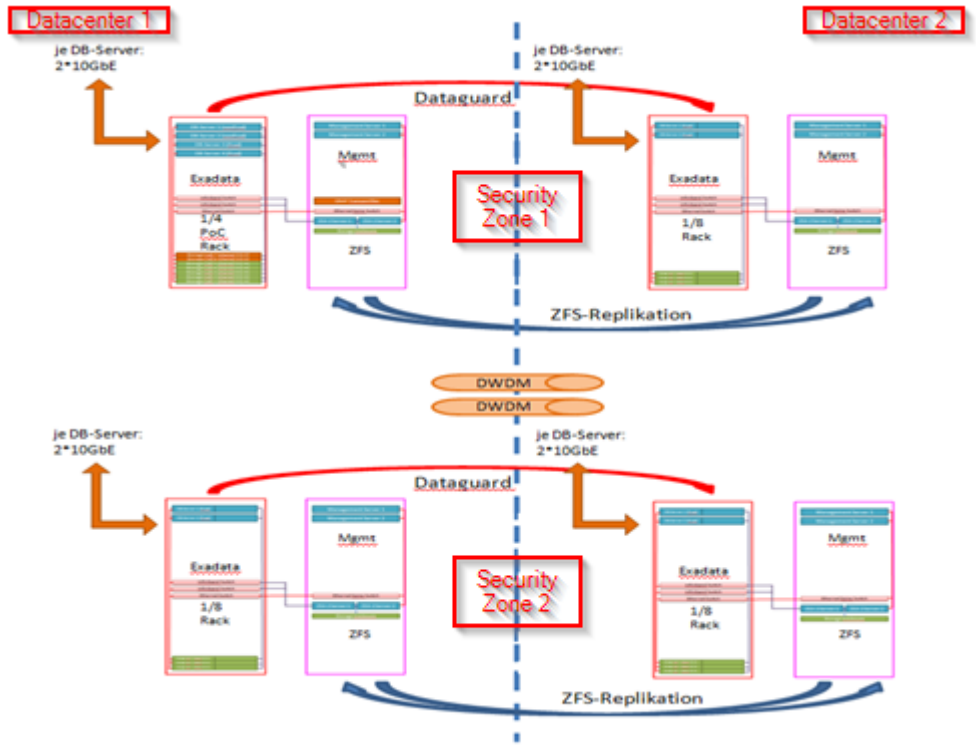

Abb. 5: Grobübersicht Architektur Konsolidierungsplattform Oracle Exadata

Nach dem Grobkonzept haben wir PoC-Kandidaten ausgewählt, die eine gewisse Herausforderung im Datenbank-Umfeld darstellen: sowohl mit Blick auf die Komplexität beim Nutzen von Features der Oracle-DB wie auch das anspruchsvolle IO-Verhalten. Das sollte sicherstellen, dass die Plattform zum einen aus technischer Sicht die allermeisten Ansprüchen meistern wird. Desweiteren wollten wir mit diesen in der Telekom-IT bekannten Kandidaten auch das „Eis brechen“ für zukünftige Migrationen. Getreu dem Motto: „Wenn bei denen der PoC erfolgreich war, dann haben wir erst recht keine Probleme“.

Die PoC Phase
Genehmigungen und/oder Bestellungen neuer Systeme dauern in oder zwischen Großkonzernen einige Zeit - zumal wenn die Kosten höher als üblich liegen.

Die Beschaffung der Exadata lief auf „Buy-to-Try“-Basis. Dafür beschrieb der erste Schritt – abgestimmt mit Oracle – den Proof of Concept (POC) und all seinen Erfüllungskriterien. Da zu den ausgewählten Systemen schon viele Performance-Kennzahlen vorliegen, war dieser Schritt einfach. Der POC galt als bestanden, wenn die Performance der ausgewählten Tests mindestens genauso gut ist wie auf der bestehenden Performance-Testplattform und dazu weniger CPUs benötigt. Ferner gab es noch festgelegte Kriterien für die Einsparung beim Storage und für die betrieblichen Tests.
Da die technischen Voraussetzungen im Rechenzentrum vorhanden waren (Strom, Netzwerk, IP-Adressen, …), war der physische Aufbau des PoC-Exadata-Systems in 2 Tagen erledigt. Ferner wurde noch ein Management-Server für den „Oracle Enterprice Manager“ (OEM) bereitgestellt. Ebenso ging später der ZFS-Filer für die Backup-Anforderungen schnell in Betrieb.

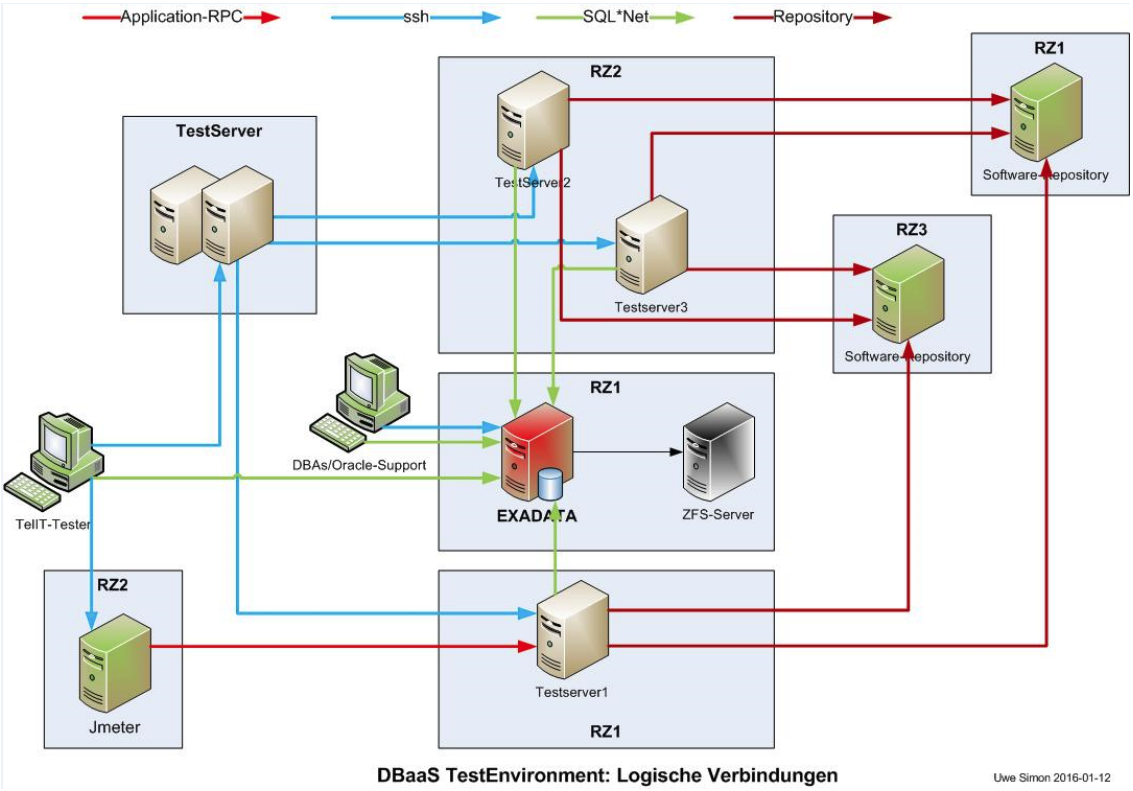

Abb. 6: Grobübersicht PoC-Aufbau

Die Installation der Umgebung für den POC benötigte im ersten Schritt 4 Oracle-Virtual-Maschines (OVM), die mit dem „Oracle Exadata Deployment Assistenten“ (OEDA) konfiguriert wurden. In diesen 4 OVMs wollten wir dann die 16 Datenbanken unseres POC-Environments installieren. Als Oracle RDBMS-Release kam Anfang 2016 noch 11.2.0.4 zum Einsatz, da die Quellsysteme zu dem Zeitpunkt noch unter 11g liefen. Mit der OEDA-Version von Anfang 2016 konnten wir die OVM noch nicht komplett in der GUI konfigurieren, da VLANs noch nicht unterstützt wurde.

Wie eigentlich bei allen Projekten, bei denen sich die Architektur großer Datenbanksysteme ändert, verursacht die Datenmigration am meisten „Schweiß“. Besonders, wenn man später in Produktion mit möglichst wenigen, kurzen Auszeiten auskommen möchte.
Solange man auf der gleichen CPU-Plattform bleibt, ist das bei Oracle ja alles mit RMAN/Dataguard/Physical-Standby etc. gut und einfach handhabbar. Sehr große Datenbanken (x-TB und hohe Transaktionsraten) verschiebt man allerdings auch damit nicht von einem ins andere Rechenzentrum. Ein einfacher NAS-Filer und eine Spedition sind hier manchmal immer noch der schnellste, kostengünstigste Weg.
Beim Wechsel der CPU macht einem Oracle das Leben nicht gerade einfach. RMAN kann leider nur Datenbanken automatisch von einer CPU-Plattform auf eine andere verschieben, solange die Endianess der beiden CPUs identisch ist. Das gilt nicht für Redologs, so dass Dataguard nicht in Frage kommt. Diese Konstellation haben wir im PoC aber nicht (Power7/8 zu Intel x86-64Bit). Somit bleibt nur der mühsame Weg über „Transportable Tablespaces“ – mit den entsprechenden Nacharbeiten (zusätzliche Skripts für Accounts, Grants, Synonyms auf Quelle generieren, usw.). Im Rahmen einer Konsolidierung von hunderten Datenbanken muss das aber möglichst ohne manuelle Aktivitäten und natürlich fehlerfrei laufen. Um kurze Auszeiten zu erreichen, muss später die produktive Migration über „Logical-Standbys“ erfolgen (besonders bei den großen DBs).
Die Datenmigration über Transportable Tablespaces hat für den Performancetest noch den „positiven Effekt“ möglichst vergleichbarer Ergebnisse. Da wir eine physische Kopie der Daten haben, entfallen - aus Sicht der Exadata - positive Effekte durch eine Reorganisation der Tabellen bzw. Indizes (bei Export/Import oder anderen Datentransferlösungen).

Bei dem für den Performancetest ausgewählten CRM-System haben sich in der Vergangenheit Änderungen an der Systemarchitektur mehr oder weniger deutlich bemerkbar gemacht. Besonders Änderungen an der Storage-Architektur haben sehr starke Einflüsse. Das ist hauptsächlich darauf zurückzuführen, dass man pro Terrabyte Tablespace immer weniger Harddisks benötigt. Gerade bei Random-Access-Zugriffen macht sich das deutlich bemerkbar. Die Anzahl der Random-Access-Zugriffe pro Disk hängt direkt von deren Umdrehungszahl ab (15000, 10000, 7200 RPM). Dies ergibt Werte zwischen 200 und 400 Random-Access-IOs/Sekunde je Disk. Eine Exadata in der HighCapacity-Konfiguration hat pro Full-Rack 168 Disks mit 7200RPM, lässt maximal ca. 32000 Random-Access-IOs/Sekunde von der Disk zu. In unserem ausgewählten CRM-System haben wir aktuell mittags in einer Datenbank schon rund 20000 Random-Access-IOs von Disks. Kritisch ist dies besonders dann, wenn auf Daten zugegriffen wird, die Wochen/Monate nicht in Benutzung waren und somit auch nicht gecached sind. Auf der bestehenden Plattform musste der Einsatz von SSDs die geringere Random-Access-IO-Rate beim Wechsel der Harddisks im Storage von 15000 RPM-Disks auf größere 10000 RPM-Disks ausgleichen.

Die Performancetests gliedern sich in

Datenbank-Funktionalität	**Fachliche Funktionalität**
Fullscan	Laden von externen Applikationscaches
RandomAccess	Kundendatenzugriff (Callcenter, Webportale, …)
RandomAccess + Fullscan	Normaler Betrieb
Fullscan + DB-Link	Laden aktueller Kundenstammdaten zur Rechnungsschreibung

Um möglichst vergleichbare Messwerte zu erhalten, fanden die ausgewählten Performancetests als Erstes auf der produktionsnahen Performance-Testumgebung statt. In der bestehenden Umgebung sind schon die kritischen Tabellen auf SSDs gespeichert. Für den ersten Test wurden diese kritischen Tabellen 1:1 auf Extreme-Flash-Storage der Exadata übertragen. In einem zweiten Testdurchlauf haben wir die Tabellen auf die „normalen Disks“ (High-Capacity-Storage) abgelegt und einen

entsprechenden Anteil des FlexCaches der Exadata konfiguriert. So hatten wir einen Vergleich, inwieweit wir mit dem Cache die Performance der langsameren Harddisks ausgleichen können.

Für den Kompressionstest haben wir die großen Objekte ausgewählt. Hier belegen wenige Tabellen den Großteil des Storage. Generell gilt ja, „was man mit den großen Objekten nicht erreicht, kann man mit den kleinen Objekten nicht mehr herausholen“. Da diese ausgewählten großen Tabellen sehr viele IDs (aus Sequences) und Zeitstempel haben (und somit die Zeilen nicht sehr ähnlich sind), war unsere Erwartung an die Kompressionsfaktoren von HCC deutlich niedriger, als die Oracle-Werbung suggeriert.

Das beste System ist „nur halb so gut“, wenn es sich nicht in das bestehende Environment integrieren lässt. Betriebliche Tests sollten das überprüfen. Hier sind die wesentlichen Punkte „Datensicherung“, „Einbinden in die bestehende Betriebsüberwachung“, „Bereitstellen von OVMs und Datenbanken“, „Messen der Auslastung“, Planen der Kapazitäten“.

Die Datensicherung erfolgt per RMAN auf Disk auf einen ZFS-Filer, der direkt über Infiniband an die Exadata angeschlossen ist. Das ermöglicht schnellen Backup und Restore gerade von großen Datenbanken.

Das PoC-Ergebnis

Wie erwartet hat sich bei der Exadata die beste Performance bei „datawarehouse-like“-Abfragen gezeigt. Hier kann das System alle seine Vorteile ausspielen. Bei Applikationen mit Einzelsatzverarbeitung kann aber auch eine Exadata die Physik nicht überlisten (eine 8TB Harddisk mit 7200 RPM kann nicht beliebig viele Random-Access-Zugriffe/Sekunde). ExtremeFlash und FlexCache können diese Nachteile beim Random-Access-IO aber kompensieren. ExtremeFlash-Storage steigerte die Performance zwischen 220 und 280 Prozent. FlexCache verlangsamte die Prozesse beim ersten Starten deutlich (Daten kommen direkt von der Disk). Je länger der Cache „warmlief“ , umso schneller wurden sie dann, um nach einigen Minuten dann fast an die ExtremFlash-Konfiguration heranzukommen.

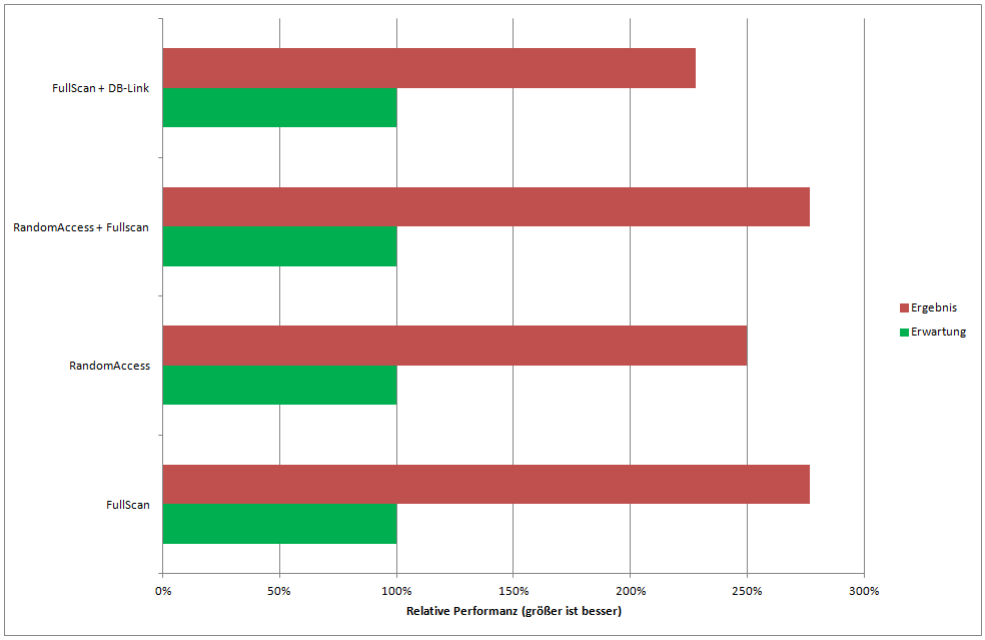

Abb.7: Zusammenfassung der Ergebnisse der Performancetests

Mögliche Storage-Einsparungen durch Hybrid Columnar Compression (HCC) hängen wesentlich von den Dateninhalten und den Zugriffen auf diese Daten ab. Hier verspricht die „Werbung“ deutlich mehr, als in unserem Fall möglich ist. Die „guten Werte“ lassen sich hauptsächlich bei „DWH-like“ Datenbanken erreichen. Bei einer Konsolidierungsstrategie auf Exadata wird es immer einige große Datenbanken geben, die von einer Kompression profitieren. Die Zahl solcher Datenbanken ist aber eher klein und da lassen sich kaum kompressionswürdige Daten finden bzw. der Aufwand dies auszunutzen steht in keinem Verhältnis zur potentiellen Einsparung. In unseren Testdatenbanken ließen sich nur rund sieben Prozent der Daten komprimieren. Die Kompressionsrate lag bei einem Faktor von 4,8. Somit ergab sich eine Einsparung von nur ca. 5 Prozent. Auch bei anderen Datenbanken zeigten sich entsprechende Faktoren bei komprimierbaren Tabellen. Generell sollte man bei der Komprimierung im Zuge von Datenbank-Konsolidierungen nicht zu optimistisch sein.

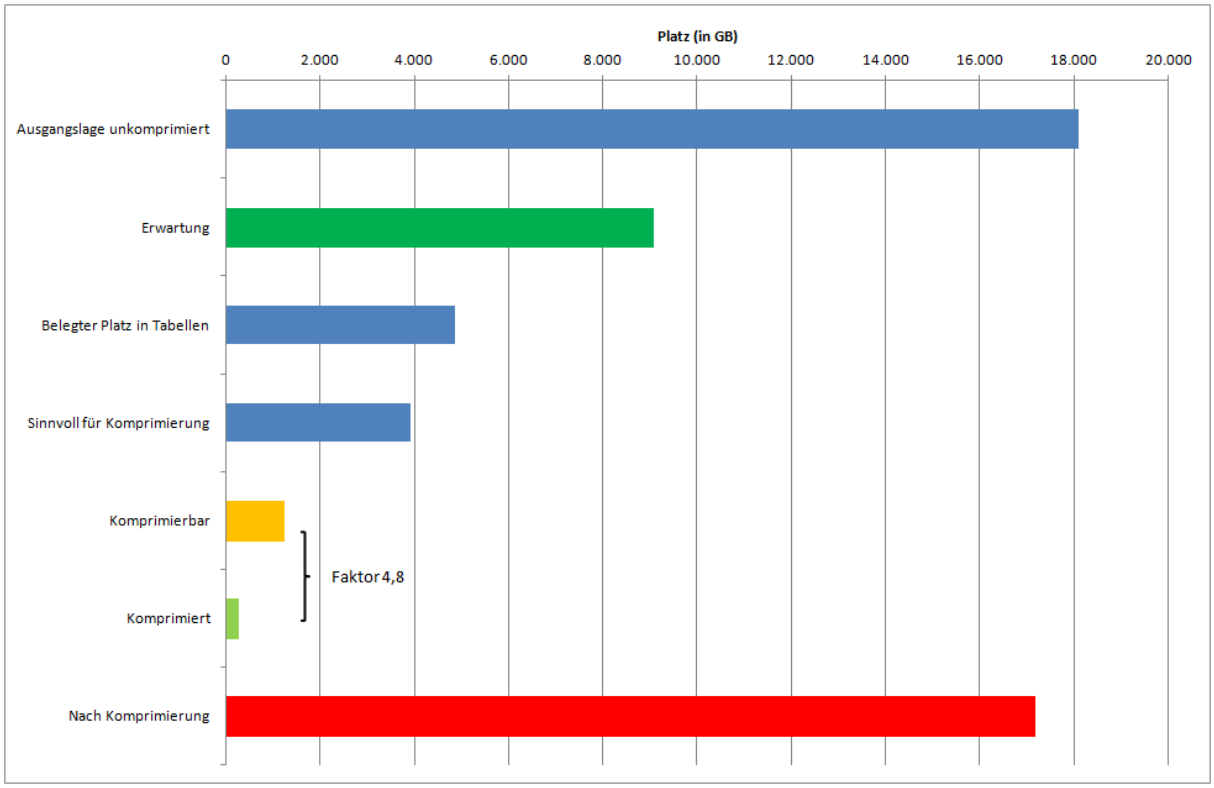

Abb.8: Zusammenfassung der Ergebnisse der Kompressionstest

Eine „rechnerische“ Ersparnis beim Storage im Vergleich zur bestehenden Infrastruktur ergibt sich aber doch: im „current state“ werden sowohl OS, Software als auch Datenbanken auf SAN-Storage abgelegt und entsprechend abgerechnet. Auf der Exadata liegen OS, Software etc. auf den lokalen Platten der DB-Server und somit steht die Kapazität der Storage-Nodes komplett für die Datenbanken zur Verfügung. Nach unserem neuen Verrechnungsmodell wird hier nur noch die allokierte Menge abgerechnet („pay what you use“).

Der Grad einer möglichen Reduktion bei den CPU-Cores hängt wesentlich von den Datenbankinhalten und deren Nutzung ab. Die Reduktion ergibt sich – prinzipbedingt - nur wenn entsprechend viele SQL-Abfragen vom Query-Offloading (Smart-Scan) der Storage-Server profitieren. Dies ist bei vielen relativ kleinen Datenbanken oder den großen CRM-Systemen meist nicht der Fall. Hier darf man auch nicht zu optimistisch herangehen.

Für den Betrieb wird eine einheitliche Konfiguration der Exadata-Systeme, der OVMs und der Datenbanken angestrebt. Nur so kann man einen hohen Automatisierungsgrad erreichen. Mit sehr wenigen Konfigurations-Parametern müssen sich möglichst viele Datenbanken „out of the box“ komplett „mit einem Click“ bereitstellen lassen können (OVM, Oracle-Software, DB-Instanz, RMAN-Backup, …). Jede Abweichung in den Konfigurationen verursacht manuelle Eingriffe.

Unsere Erwartungshaltung für den Produktionsbetrieb ist, dass bei steigender Auslastung der Exadata die Performance der einzelnen Applikationen zwar nachlassen wird, aber weiterhin über dem Niveau der bestehenden Plattformen bleibt. Unser Ziel ist ja, die Exadata-Systeme - bei gleicher Performance wie bisher - möglichst gut auszulasten. Um dies zu erreichen, müssen wir im laufenden Betrieb ggf. mit weiteren ExtremFlash-Modulen - zu Lasten des Business-Cases - nachrüsten.
Aus Sicht des „Technikers“ ist es immer gut, wenn man „Reserven“ hat und damit die Performance noch steigern kann. Das reduziert das technische Risiko. Hier muss man aber immer zwischen Business Case und Performance abwägen, es gibt eben nichts „umsonst“.

Die nächsten Schritte
Nach Abschluss des PoC mit positivem Ergebnis und einem ebenfalls sehr positiven Gesamt-Business-Case hat die Telekom-IT entschieden, mit dieser Plattform für Oracle Datenbanken produktiv zu gehen.
Es wurden die Betriebsabläufe weiter optimiert und die Bestell-Prozesse entsprechend angepasst. Die Migrationsverfahren, um Datenbanken von AIX/Solaris zu migrieren, haben wir weiter verfeinert.
Das OVM-Handling ist noch nicht so, wie man es sich vorstellen könnte. So deckt der OEM noch nicht alles ab, was mittels Virtualisierung auf der Exadata möglich ist. U.a. das Handling von mehreren VLANs an einem DB-Server ist noch manuell. Wir stehen dazu noch im Kontakt mit der Oracle-Entwicklung, um hier eine Lösung zu finden.
Der Plan ist nun, Anfang 2017 die komplette Plattform über zwei Data-Center-Standorte aufzubauen. An jedem Standort werden wir dann zwei physisch getrennte Security-Zonen aufbauen, deren Datenbanken bei Anforderung über Dataguard an den zweiten Standort repliziert werden.

Kontaktadresse:
Gregor Büchner
T-Systems International GmbH
Wolbeckerstr. 268
D-48155 Münster
Telefon: +49 (0) 251-3977 2814
E-Mail gregor.buechner@t-systems.com
Internet: www.t-systems.com

Uwe Simon
T-Systems International GmbH
Landgrabenweg 151
D-53227 Bonn
Telefon: +49 (0) 228-181 36760
E-Mail uwe.simon@t-systems.com
Internet: www.t-systems.com

Joachim Dietsch
ORACLE Deutschland B.V. & Co. KG
Robert-Bosch-Strasse 5
D-63303 Dreieich
Telefon: +49 (0) 6103-397 216
E-Mail joachim.dietsch@oracle.com
Internet: www.oracle.de