Mit den funktionalen Erweiterungen der Oracle-Datenbank 11.2.0.2 und dem zwischenzeitlich aktualisierten Exadata Storage Server 11.2.2.3.2 steht nun für nahezu jedes Einsatzszenario eine Lösung bereit.

Neuigkeiten über Exadata

Frank Schneede, ORACLE Deutschland B. V. & Co. KG

Dieser Artikel gibt einen Überblick über die neuen Funktionen, die exklusiv auf der Exadata-Plattform zur Verfügung stehen. Der Schwerpunkt liegt auf Exadata als Konsolidierungsplattform. Darüber hinaus werden die neuen Funktionen kurz vorgestellt, die in der aktuellen Version des Exadata Storage Servers und in der Betriebssystemalternative Solaris 11 Express enthalten sind.

Alle Modelle der Exadata Database Machine stellen dem Anwender außergewöhnlich großzügig bemessene Ressourcen zur Verfügung. So gibt es bereits in der kleinsten Ausbaustufe, dem X2-2 Quarter Rack, High-Performance-SAS-Festplatten und 21 TB Festplattenplatz (Rohdaten), bei einer möglichen I/O-Bandbreite von 5,4GB/Sek. und einer I/O-Leistung von bis zu 10.800 IOPS (auf Festplatten) beziehungsweise 375.000 IOPS (auf Flash). Das Topmodell, die X2-8 mit High-Performance-SAS-Festplatten, bietet sogar 100 TB Festplattenplatz (Rohdaten), eine I/O-Bandbreite von bis zu 25GB/Sek. und eine I/O-Leistung von 50.000 IOPS (auf Festplatten) beziehungsweise 1.500.000 IOPS (auf Flash). Die genannten Kennzahlen beziehen sich auf die Größen, die sich allein aus der Hardware-Konfiguration ergeben. Bei Verwendung von Komprimierungsmethoden wie Exadata Hybrid Columnar Compression (EHCC) lassen sich diese noch um Faktoren verbessern.

Systeme einer solchen Größenordnung sind auch wegen des damit verbundenen Investitionsvolumens prädestiniert für die Konsolidierung von Datenbanken. Datenbanken als Grundlage verschiedener Anwendungen haben häufig jedoch unterschiedliche Last-Charakteristika; man unterscheidet grob zwischen OLTP- und Data-Warehouse-Transaktionslasten. Um unterschiedliche Produktivdatenbanken auf einem System wie der Exadata Database Machine konsolidieren zu können, ist es wichtig, dass jede Datenbank über ein Service Level Agreement eine entsprechende Ressourcennutzung zugesichert bekommt. Die Exadata Database Machine und die Datenbank Oracle 11.2.0.2 verfügen über Mechanismen, um die Ressourcenverteilung zu optimieren.

Ressourcen geschickt verteilen

Bei einer Konsolidierungsstrategie unterscheidet man üblicherweise zwischen Server-, Storage- oder Datenbank-Konsolidierung. Es gibt aber auch bei einer einzelnen Datenbank unterschiedliche Lasten, die um die Verwendung der vorhandenen Ressourcen konkurrieren. Ein Beispiel dafür sind Berichte, die auf einer klassischen OLTP-Datenbank laufen. Je nach der gewählten Konsolidierungsstrategie sind diese so zu verteilen, dass jede Datenbank zu jeder Zeit die Ressourcen zugewiesen bekommt, die zur Erledigung der laufenden Arbeiten notwendig sind. Dies betrifft folgende Bereiche:

- CPU-Ressourcen
- I/O-Ressourcen
- Parallele Prozesse innerhalb einer Datenbank

Die Ressourcen-Zuweisung kann dabei nach Tageszeiten variieren.

Für die Verteilung der CPU-Ressourcen ist der schon länger bekannte Database Resource Manager zuständig, der die Prozesse einer Datenbankinstanz auf alle vorhandenen CPUs des Datenbankservers verteilt. Bei einer Exadata Database Machine X2-2 stehen dafür auf einem RAC-Knoten zwölf Prozessorkerne zur Verfügung. Eine Beschränkung der Anzahl der Prozesse, die zu einem bestimmten Zeitpunkt laufen, ist durch das Feature „Instance Caging“ möglich. Es steht seit der Datenbankversion 11.2.0.1 auch für Nicht-Exadata-Systeme zur Verfügung. Der Initialisierungsparameter „cpu_count“ legt die maximale Anzahl der CPUs fest, die eine Datenbankinstanz erhält. Voraussetzung ist die Aktivierung des Database Resource Managers durch den Initialisierungsparameter „resource_manager_plan = <plan_name>“.

Bei der Ressourcenverteilung gilt es einerseits, jederzeit die notwendigen Ressourcen zuzuweisen, und andererseits, zu Zeiten geringer Systembelastung nicht mehr Ressourcen zu vergeben, als der entsprechenden Aufgabe zustehen und im Rahmen der Leistungsverrechnung ermittelt werden. Die Direktive „max_utilization_limit“ in der Database-Resource-Manager-API „dbms_resource_manager“ steuert die Beschränkung der CPU-Nutzung. Seit Version 11.2.0.2 gibt es weitere Direktiven, um beispielsweise die Zeit oder die Ausführungspriorität einer Abfrage zurückzusetzen oder die Verwendung von Parallel-Server-Prozessen zu begrenzen.

Hand in Hand mit dem Database Resource Manager arbeitet ein I/O Resource Manager, der bereits in der ersten Version des Exadata Storage Servers enthalten war. Im Gegensatz zum Database Resource Manager, der nur die Ressourcen einer Datenbankinstanz verwaltet, arbeitet der I/O Resource Manager für den gesamten Shared Storage. Das bedeutet, dass man I/O-Anfragen aller auf einen Shared Storage zugreifenden Datenbanken nach flexiblen Kriterien priorisieren kann. In der aktuellen Exadata-Version ist analog zum Database Resource Mana-

ger die Möglichkeit hinzugekommen, auch bei geringer Systembelastung die I/O-Ressourcen zu beschränken.

Seit der Version 11.2.0.2 besitzt der I/O Resource Manager einen Automatismus, über den gesteuert werden kann, wie die I/O-Nutzung optimiert werden soll. Der Parameter „objective" gibt die Optimierungsstrategie für den I/O Resource Manager an und kann folgende Werte annehmen:

- *low_latency*
  Für OLTP-Systeme, die eine besonders geringe Latenz für den Festplattenzugriff benötigen
- *balanced*
  Mit dieser Einstellung wird eine Balance zwischen Latenz und I/O-Durchsatz erreicht
- *high_throughput*
  Für Data-Warehouse-Systeme, die einen sehr hohen Durchsatz benötigen
- *auto*
  Der I/O Resource Manager ermittelt aus aktueller Last und dem aktivierten Resource-Management-Plan die bestmögliche Optimierungsstrategie
- *off*
  Abschalten der I/O-Ressourcen-Zuweisung

Die genannten Erweiterungen in Database und I/O Resource Manager tragen der gestiegenen Bedeutung der Exadata Database Machine als Konsolidierungsplattform Rechnung. Gerade Dienstleister, die auf ihren Systemen verschiedene Kundendatenbanken betreiben wollen und Service Level Agreements mit ihren Kunden vereinbaren, profitieren von einer möglichst granularen Ressourcen-Zuweisung, die sowohl Ober- als auch Untergrenzen berücksichtigt. Eine auf der Ressourcen-Verwendung basierende Auswertung zur Abrechnung (Charge-Back) steht jedoch noch aus. Sowohl für den Database Resource Manager als auch für den I/O Resource Manager steht aber mit dem Oracle Enterprise Manager Grid Control (OEMGC) ein Werkzeug zur Verfügung, mit dem eine grafische Überwachung der Systemaktivitäten möglich ist.

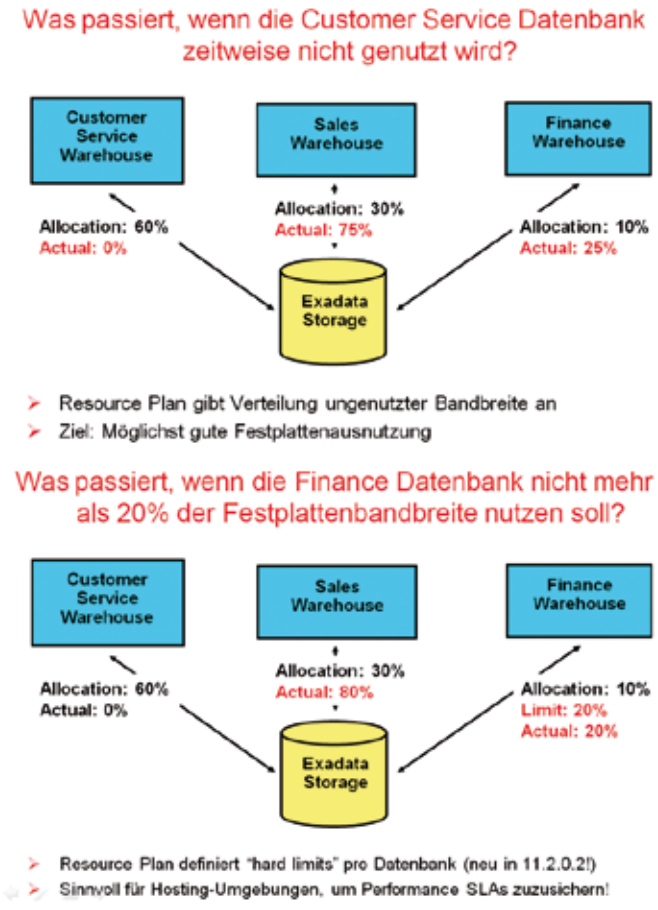

Abbildung 1: „Hard Limits" im I/O Resource Manager

Solaris 11 Express

Bereits mit der Vorstellung der Exadata wurde die Möglichkeit angekündigt, auf den Datenbankservern das Betriebssystem Solaris 11 Express einsetzen zu können. Diese Alternative steht nun zur Verfügung. Kunden können bei der Einrichtung der Exadata Database Machine zwischen Oracle Enterprise Linux und Solaris 11 Express entscheiden. Nach dem initialen Setup ist eine Änderung des Betriebssystems nur durch ein vollständiges Neuaufsetzen der Maschine (Re-Imaging) möglich.

Solaris 11 Express ist ein vollständig unterstütztes Solaris-Betriebssystem für X86-Plattformen, das auf der Exadata Database Machine jedoch nicht über den vollen Funktionsumfang (zum Beispiel keine Virtualisierungsmöglichkeiten wie Solaris-Zonen) verfügt. Trotzdem ist Solaris 11 Express besonders für Exadata-Kunden interessant, die strategisch auf die Plattform „Solaris" setzen.

Die Architektur der Exadata Database Machine ist ohne „single point of failure" ausgelegt – Komponenten wie Festplatten oder die Stromversorgung sind während des laufenden Betriebs austauschbar. Automatismen erkennen und melden mögliche Problemsituationen noch vor deren Eintreten. Auf Betriebssystemebene steht nun mit dem sogenannten „Predictive Self Healing" eine Funktion zur Verfügung, die dazu dient, die Servicezeiten der Exadata Database Machine zu verbessern. Darunter ist in erster Linie ein Diagnosewerkzeug zu verstehen, das die Zustandsmeldungen des Systems für Hard- und Software sammelt und auswertet. Als Ergebnis können automatisch Maßnahmen zur Wiederherstellung oder zur Weitermeldung ergriffen werden. Der Oracle Solaris Fault Manager koordiniert die Problemanalyse und sorgt dafür, dass aussagekräftige Fehlermeldungen sowie Recovery-Maßnahmen erfolgen. Der Solaris Service Manager startet die fehlerbedingt abgebrochenen Dienste unter Berücksichtigung der Abhängigkeiten neu.

Das neue Tool Oracle Solaris DTrace analysiert und tunt Anwendungen auf der Exadata Database Machine weitgehend ohne Beeinträchtigung des Produktionsbetriebs. Es liefert Trace-Informationen aus dem Betriebssystem, um Engpässe zu erkennen und zu beheben.

Die Installation und Wartung des Solaris-11-Express-Betriebssystems erfolgt durch das neue Image Packaging System (IPS). Es bietet ein Framework, das den gesamten Software-Lebenszyklus abdeckt, also Installation, Upgrade und Deinstallation von Packages. Bei Verwendung des ZFS-Filesystems und des Boot-Environments sind System-Upgrades vollständig abgesichert.

Die außergewöhnliche Performance der Exadata Database Machine beruht unter anderem auf den Erweiterungen des „zero-loss zero-copy"-InfiniBand-Protokolls RDS V3, über das die Storage-Server mit den Datenbank-Servern sowie die Datenbank-Server untereinander kommunizieren. Solaris 11 Express unterstützt dieses Protokoll und liefert dadurch die erwartete I/O-Bandbreite bei geringer Latenz. Exadata basiert auf der sogenannten „NUMA-Architektur" (Non-Uniform Memory Access I/O), in der jede CPU beziehungsweise jede Gruppe von CPUs über exklusiv zugeordneten physikalischen Speicher und I/O-Devices verfügt. Solaris 11 Express unterstützt diese Architektur, indem es Betriebssystem-Ressourcen (Interrupts, Kernel Threads und Speicher) unter Berücksichtigung der physikalischen System-Topologie und der aktuellen Systemlast den physikalischen Ressourcen zuordnet.

Intimate Shared Memory (ISM) verbessert im Solaris-Betriebssystem die Performance innerhalb des Softwarestacks der Oracle-Datenbank auf Systemen mit großem Hauptspeicher. Das Anlegen und Locking-Verhalten von ISM einer festen Größe ist in Solaris 11 Express verbessert worden, sodass die Startup-Performance der Oracle-Datenbank um bis zu Faktor acht gesteigert werden konnte. Da die Größe der SGA der Oracle-Datenbank mittlerweile dynamisch angepasst werden kann, ist auch für den ISM eine Anpassung während der Laufzeit (Dynamic Intimate Shared Memory) möglich.

Memory Placement Option

Weitere Optimierungen in der Speicherverwaltung von Solaris 11 Express wie die Memory Placement Option (MPO) tragen ebenfalls zu Performance-Verbesserungen bei. Grundidee der MPO ist es, Speichersegmente möglichst nahe an den Prozessoren zu platzieren und dadurch Zugriffszeit einzusparen. Durch Einsatz des Multiple Page-Size Support werden Speicherzugriffe über den Translation Lookaside Buffer (TLB) für die Zuordnung der logischen zur physikalischen Adresse eines Speichersegments optimiert. Dieses Vorgehen ist besonders dann vorteilhaft, wenn man mit großen Datenmengen arbeitet.

Solaris 11 Express bietet ein hohes Maß an Sicherheit, da es nur mit dem notwendigen Minimum an Services installiert wird. Das aus dem Bereich der Datenbank-Security bekannte Prinzip des Least Privilege ist über ein Rollenkonzept für administrative Aufgaben im Betriebssystem umgesetzt. Das Solaris-Audit-Feature protokolliert Aktivitäten im System auf einer granularen Ebene. Die Einhaltung von Compliance-Regeln ist sichergestellt, weil Betriebssystembenutzer klar definierte Rollen und Zugriffsrechte besitzen.

Exadata Storage Server Software 11.2.2.3.2

Neben den Exadata-spezifischen Neuerungen in der Oracle-Datenbank 11.2.0.2 und Solaris 11 Express sind auch in die Softwarelösung des Exadata Storage Servers einige wichtige Erweiterungen eingeflossen. Neben Aktualisierungen der Firmware für InfiniBand-Komponenten, Server ILOM und BIOS sowie Festplatten-Controller sind Verbesserungen in der Überwachung der Systemsicherheit durch neue Events, die die Exadata-Zelle sendet, in der neuen Version enthalten. Auf zwei neue Funktionen wird detaillierter eingegangen: Exadata Secure Erase und Optimized Smart Scan.

Kunden, die ihre Datenbank auf einer Exadata nur temporär betreiben, haben aus Datensicherheitsgründen ein großes Interesse daran, dass alle Daten rückstandslos wieder gelöscht werden, sobald die Maschine nicht mehr im Einsatz ist. Dieser Anforderung trägt Exadata Secure Erase Rechnung. Es stehen drei verschiedene Modi zur Verfügung, die in ein, drei oder sieben Iterationen die auf den Cell Disks oder Grid Disks befindlichen Daten mit bewährten Algorithmen durch unterschiedliche Zeichendaten überschreiben. Der Zeitbedarf je Platte ist unter Umständen erheblich, wie Tabelle 1 zeigt.

Laufwerkstyp	1pass	3pass	7pass
600 GB	1 Std.	3 Std.	7 Std.
2 TB	5 Std.	15 Std.	35 Std.
Flash (22,875GB)	n/a	n/a	21 Min.

Tabelle 1: Zeitbedarf für das Überschreiben von Datenträgern

Newsticker

Loïc le Guisquet sagt Konferenz ab

Loïc le Guisquet, Oracle Executive Vice President EMEA, der als Keynote-Speaker auf der DOAG 2011 Konferenz erwartet wurde, hat seine Teilnahme abgesagt. Aufgrund eines internen Management-Meetings könne er an der Konferenz nicht teilnehmen, hieß es. Es war der dritte Versuch der DOAG, den EMEA-Chef für die Konferenz zu gewinnen. „Ich denke, das ist eine vertane Chance von Oracle den Kunden gegenüber", so Dr. Dietmar Neugebauer, Vorstandsvorsitzender der DOAG. „Nach der schnellen Zusage im Juni 2011 haben wir uns sehr gefreut, Loïc le Guisquet auf unserer Konferenz begrüßen zu können. Umso mehr waren wir über die nur sehr kurz gehaltene Absage-E-Mail enttäuscht." Durch die Vielfalt des Programms kann die Absage des EMEA-Chefs sicherlich mehr als ausgeglichen werden. Allerdings hätten die Teilnehmer seine Anwesenheit als wichtigen Schritt in Richtung Anwendernähe gedeutet. Ob le Guisquet im nächsten Jahr zur Konferenz kommt, kann Neugebauer nicht sagen: „Der Ball liegt bei ihm", meint er.

SQL Offloading – auch als „Smart Scan“ bezeichnet – ist eine Schlüsselfunktionalität und ein Alleinstellungsmerkmal der Exadata Database Machine. Die Verlagerung bandbreitenintensiver Lese- und Filteroperationen auf den Storage bringt eine Reduzierung des I/O-Volumens, das zum Datenbank-Knoten transportiert werden muss. Die dadurch freiwerdenden Ressourcen (Netzwerk und CPU) stehen für andere Aufgaben zur Verfügung, sodass insgesamt ein sehr hoher Durchsatz bei geringer CPU-Belastung der Datenbankserver in der Exadata Database Machine erreicht wird. In manchen Situationen kann es dazu kommen, dass die CPU im Exadata Storage Server sehr stark belastet wird und daher keine optimale Performance mehr erreicht werden kann. Der Optimized Smart Scan vermeidet diese Situationen.

Der Exadata Storage Server überwacht die CPU-Last, während auf dem Datenbank-Server das Scheduling der I/O-Operationen und das Monitoring der relevanten Waitevents (Resource Manager, Smart I/O) erfolgt. Auf Basis dieser Informationen erhält der Storage-Server eine Indikation, ob ein sogenannter „Push-Back“ – die Übertragung von ausgelagerten Aufgaben zurück an den Datenbankserver – erfolgen kann. Wenn währenddessen der Storage-Server eine CPU-Last von mehr als 90 Prozent ermittelt, wird ein kleiner Prozentsatz der Datenbankblöcke als herkömmlicher Block-I/O an den Datenbankserver gesendet, ohne vorher gefiltert, entschlüsselt oder dekomprimiert zu werden. Der Prozentsatz des herkömmlichen Block-I/O steigt schrittweise, solange die CPU-Last auf dem Storage-Server die Grenze von 90 Prozent überschreitet. Herkömmlicher Block-I/O kann dann bis zu 50 Prozent des gesamten I/Os ausmachen. Wenn die CPU-Last wieder unter 90 Prozent sinkt, reduziert sich der Prozentsatz des Block-I/Os wieder schrittweise. Mit diesem Verfahren kann die Performance für sehr CPU-intensive Abfragen um bis zu 30 Prozent steigen, weil die CPU-Last zwischen Storage- und Datenbank-Servern angeglichen wird.

Eine Überwachung der Optimized-Smart-Scan-Aktivitäten ist auf der Ebene des Storage-Servers möglich, der dafür mehrere Statistiken pflegt:

- CPU-Last und Push-Back-Volumen der letzten 30 Minuten
- Anzahl der 1-MB-Blöcke, die für Push-Back vorgesehen waren
- Anzahl der Blöcke mit Push-Back zum Datenbankserver
- Der „total cpu passthru output IO size“ zeigt die Menge des gelieferten I/Os in KB an

Auf Datenbank-Ebene zeigt die Session-basierte Statistik „cell physical IO bytes pushed back due to excessive CPU on cell“ an, welche Datenmenge ohne Smart Scan zurück zum Datenbank-Server gegeben wurde.

Fazit

Basierend auf den Erfahrungen aus zahlreichen Kundenprojekten ist die Software der Exadata-Produktfamilie optimiert worden. Sicherheits- und Performanceverbesserungen wurden ebenso umgesetzt wie die Anforderungen, die sich aus einem Konsolidierungsbetrieb mit der Exadata Database Machine ergeben. Nicht in Vergessenheit geraten darf an dieser Stelle die Oracle-Datenbank 11*g* R2, die im Zusammenwirken mit der Exadata Database Machine eine gute Lösung ergibt.

Weiterführende Informationen

http://www.oracle.com/us/products/database/exadata/index.html
http://www.oracle.com/technetwork/server-storage/solaris11/overview/index.html

Frank Schneede
ORACLE Deutschland B. V. & Co. KG
frank.schneede@oracle.com