Storage-Optimierung mit Oracle 11*g* R2

Afred Schlaucher, ORACLE Deutschland B.V. & Co. KG

Das Thema „Daten" ist nach wie vor einer der herausragenden Aspekte in der IT. Durch neue Techniken ist das Sammeln von Daten einerseits leichter geworden, gleichzeitig ist es Bestandteil vieler neuer Geschäftsideen und -prozesse.

Es klingt nach einer allgemeinen Aussage: Die entstehenden Datenmengen wachsen schneller als vor Jahren vermutet. Allerdings entsteht bei Verantwortlichen ein neuer, bewussterer Umgang mit dem Phänomen „Datenwachstum", denn es bedeutet vor allem mehr Kosten.

Mit dem Aufkommen der Client/Server-Systeme zu Beginn der 1990er Jahre haben sich schnell auch in der Hardware-Systemlandschaft neue Standards herausgebildet und mittlerweile zu einem festen Schema von Hardware-Architekturen und ganzen Hardware-Parks geführt. Dabei verselbstständigten sich die Hardware-Segmente „Server", „Storage" und „Netzwerk", was in der Folge oft zu organisatorischen Strukturen in den IT-Abteilungen führt:

- Um Synergien zu nutzen, konzentrierte man Speicherplattensysteme in zentral verwalteten SAN-Landschaften
- Server-Maschinen wurden zu virtuellen Rechnerverbünden gekoppelt
- Netzwerkverbindungen wurden unternehmensweit zu Firmen-internen „Autobahnen" standardisiert

Diese Entwicklung verstellt zunehmend den Blick auf die neuen Herausforderungen, die gerade große Datenmengen mit sich bringen. Ein spezialisierter und hochgezüchteter Umgang mit Storage, Servern und Netzen führt aufgrund der wachsenden Herausforderungen von hohen Datenmengen nicht nur zu gewaltigen Kosten, sondern vor allem auch zu einer Vernachlässigung von typischen Anforderungen einer Datenbank an Storage-Systeme.

Das Konkurrieren der Datenbank mit beliebig vielen Anwendungen im Unternehmen um Storage, Netzleitungen und CPU-Leistung ist ein Aspekt. Die fehlende Datenbank-spezifische Konfiguration dieser Ressourcen ist ein weiterer wichtiger Punkt, der oft zu kurz kommt.

Administratoren wurden zudem mit einer Fülle zusätzlicher Aufgaben des Storage-Designs wie Striping, Mirroring oder einfach nur der Verteilung der Datenbank-Files auf physische Platten behelligt. Konsequenzen hat dies neben Kosten auch für Backup-Konzepte und die Datensicherheit, wenn man etwa an Block-Corruption denkt. Neue Entwicklungen wie ASM und vor allem auch die der Exadata Database Machine durchkreuzen diese Entwicklung. Sie fordern zum Teil ein Umdenken bei der Gestaltung organisatorischer Strukturen in den IT-Abteilungen.

ASM minimiert Verwaltungsaufwand

Bereits mit dem Release 10 führte Oracle das Automatic Storage Management (ASM) ein. Damit definiert der Datenbank-Administrator (und nicht die Storage-Abteilung) Disk Groups als ein Set physischer Storage-Einheiten (Platten oder partitionierte Speicherbereiche von Platten). In diesen Disk Groups werden die Datenbankdateien ohne eine genaue Festlegung auf physische Platten beziehungsweise Partitionen platziert. Die ASM-Software übernimmt die plattenübergreifende, gleichmäßige Verteilung der Daten. Sie garantiert damit nicht nur das übliche Striping, sondern verhindert durch eine permanente Kontrolle auch sich anbahnende Hotspots. Außerdem wird gespiegelt (2- oder 3-fach). Kommen zusätzliche Platten in den Verbund, verteilt ASM die Daten vollautomatisch. ASM bietet folgende Vorteile:

- Durch Datenverteilung ist optimaler IO-Durchsatz garantiert
- ASM sorgt für eine einheitliche Storage-Bereitstellung für mehrere Datenbanken. Die Administration muss nur einmal erfolgen
- Teure Volume-Manager-Software ist nicht notwendig, da ASM kostenfreier Bestandteil der Datenbank ist

Erfahrungen zeigen bei einem Wechsel von traditionellem Storage auf ASM eine Performance-Steigerung von bis zu 25 Prozent.

Intelligent Data Placement

In der Datenbank 11*g* R2 hat Oracle sein ASM um sinnvolle Features erweitert. „Intelligent Data Placement" speichert häufig genutzte Daten transparent für den Benutzer automatisch auf

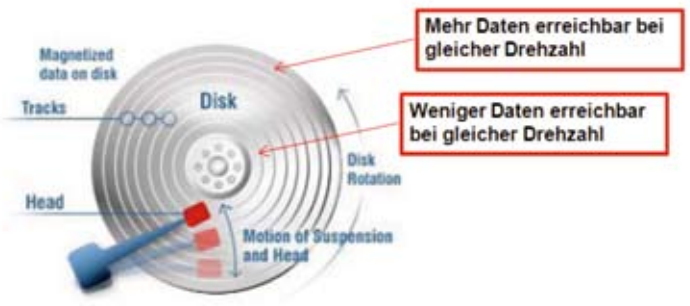

Abbildung 1: Intelligent Data Placement

die äußeren Tracks der Platten (siehe Abbildung 1). Je nach Bauart der Platte nehmen diese fünf- bis zehnmal mehr Daten auf. Leseoperationen können damit auch mehr Daten bei gleichem Aufwand lesen. Dieser Effekt nutzt vor allem dem großvolumigen, sequenziellen Lesen, wie es häufig bei Data-Warehouse-Systemen der Fall ist.

Systemverantwortliche planen gerne mehr Platten für ihre Storage-Systeme ein, auch wenn tatsächlich nicht so viel Volumen-Kapazität gebraucht wird. ASM beschreibt alle zur Verfügung stehende Platten gleichmäßig bei den äußeren Tracks beginnend; die inneren bleiben bei überschüssigem Plattenplatz leer.

11*g* R2 bietet zudem mit dem ASM Cluster File System (ACFS) die Möglichkeit, auch Nicht-Datenbank-Files über ASM zu verwalten. ACFS ist als offenes Filesystem bequem auch von dritter Seite her beschreibbar. Alle Vorteile des Stripings und Mirrorings gelten auch für diese Dateien. Das Überführen von Daten in die Datenbank (etwa über External Tables) ist bis zu dreimal schneller als die herkömmliche Verarbeitung. ACFS-Dateien sind zudem komprimierbar.

Optimierung des Durchsatzes durch optimale Anzahl von Platten

Dass die Plattenanzahl einen performancekritischen Faktor darstellt, ist bekannt. Aber wie bestimmt man die richtige Anzahl? Die Performance-Leistung von Storage-Systemen ist einerseits von der Menge der Storage-Zugriffe pro Sekunde für Inserts, Updates und Deletes bestimmt (IOPs, typische OLTP-Anforderung), auf der anderen Seite von der Datenmenge die eine Platte pro Sekunde bei lesenden Zugriffen liefert (MB/Sec, gerade bei DWH-Systemen ein wichtiger Indikator). Die Werte addieren sich, je mehr Platten im Einsatz sind. Je nach Hersteller liefern SAS-Platten etwa 100 IOPs und 30 MB/sec, SATA-Platten etwa 50 IOPs und 20 MB/sec (Herstellerangaben beachten). Wenn man für OLTP-Systeme von 5 IOPs pro Transaktion ausgeht, dann berechnet sich die Anzahl der Platten wie folgt:

$$\text{Anzahl Platten} = \frac{\text{Anzahl der zu errichenden Transaktionen pro Sek.} * 5}{\text{Erreichbare IOPs pro Platte}}$$

Heutige CPUs benötigen etwa 200 MB Datenvolumen, um vollständig ausgelastet zu sein. Wenn eine Platte etwa 20 MB/sec Datenvolumen im Lesezugriff liefert, dann berechnet sich die Menge der Platten für optimal versorgte DWH-Systeme mit vielen Leseoperationen:

$$\text{Anzahl Platten} = \frac{\text{Anzahl Cores} * 200}{20}$$

Das Tool ORION (Oracle I/O Calibration) misst den Plattendurchsatz einer Umgebung. Ab 11*g* R2 liegt das Tool im Bin-Verzeichnis der Oracle-Software und wird unabhängig von der Datenbank direkt von der Betriebssystem-Ebene aus gestartet. Eine Kurz-Dokumentation ist leicht im Internet zu finden.

Die Betrachtung der Performance des Gesamtsystems hängt natürlich von noch weiteren Faktoren ab, wobei sich der Fokus vom Plattenspeicher weg bewegt. So ist beispielsweise für die Größe des Hauptspeichers in GB die Anzahl der CPU-Cores mit 2 zu multiplizieren. Die Anzahl der Platten-Controller ist von Herstellerangaben über Controller-Durchsatz und dem zu leistenden Gesamtdatendurchsatz abzuleiten.

Solid State Disks

Eine sehr spannende Entwicklung stellen Flash-PCI-Karten-Speicher dar. Dahinter verbirgt sich ein spindelloser, direkt adressierbarer und persistenter Flash-Speicher, der über den entsprechenden Daten-Bus direkt durch das jeweilige Programm (in diesem Fall die Datenbank) beschrieben werden kann. Keine Platten-Controller und keine langsam drehenden Platten bremsen den Datenfluss. Flash-Speicher sind etwa 100- bis 200-mal schneller als klassische Plattenspeicher, allerdings mit 40 Dollar pro GB noch 20 bis 40 mal so teuer.

Häufig benötigte Daten werden nach wie vor über die SGA in dem sehr schnellen RAM als Cache-Daten für wiederholtes Lesen vorgehalten. Sinkt die Nutzungshäufigkeit, so verlagert die Datenbank diese Daten in den Flash-Speicher. Noch seltener genutzte Daten müssen in der Regel direkt von den Platten gelesen werden.

Den Flash-Speicher richtet man über die beiden „INIT.ORA“-Parameter „db_flash_cache_file = <filename>“ und „db_flash_cache_size=<size>“ ein. Die transparente und automatisierte Verwendung des Flash-Speichers durch die Oracle-Datenbank-Software weitet den Performance-Zugewinn auf die gesamte Datenbank aus, auch wenn deren Volumen die Größe des Flash-Speichers weit übersteigt. Flash-Speicher werden künftig eine noch wesentlich größere Rolle spielen, da ihre anwendungsgerechte Nutzung durch die Datenbank der Nutzung von Flash-Speichern in reinen Storage-Systemen überlegen ist.

Storage Protection verhindert Datenverlust

Block-Corruption ist ein weiteres Feld, in dem die Datenbank-basierte Verwaltung reinen Storage-Systemen überlegen ist. Bereits in 11*g* R1 wurde die

Abbildung 2: Unterschiedliche Speicher-Medien hinsichtlich Preis und Leistung

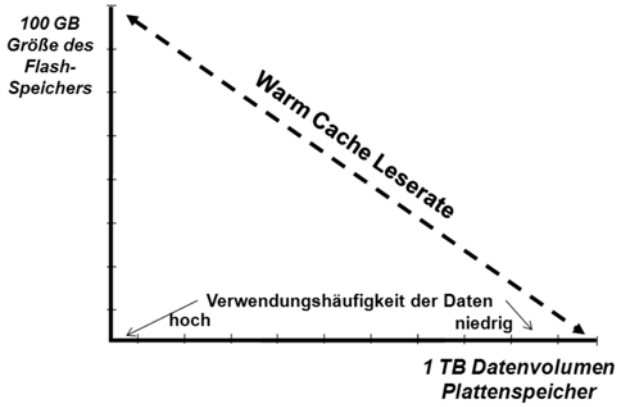

Abbildung 3: Performance-Gewinn durch Flash-Speicher

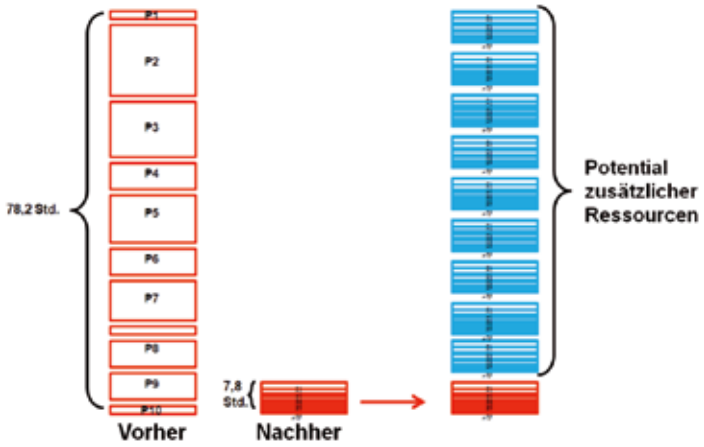

Abbildung 5: Exadata spart Ressourcen

neue Funktion „DB_LOST_WRITE_PROTECT=[NONE|TYPICAL|FULL]“ eingeführt. Diese verhindert eine Situation, in der das Storage-Subsystem meldet, dass Daten geschrieben wurden, während dies tatsächlich nicht stattfand. Mithilfe von Check-Summen und Redo-Daten validiert die Datenbank-Software die Konsistenz der gespeicherten Blöcke. In derselben Weise prüft auch RMAN die Backup-Daten, bevor er sie in den Sicherungsbestand schreibt.

Mit Exadata an die Spitze eines neuen Trends gesetzt

Mit der Exadata Database Machine adressiert Oracle konsequent die Herausforderungen von wachsenden und teuren Datenbergen und hat sich an die Spitze einer Bewegung gesetzt, die, fast unbemerkt von der breiten IT-Community, mit tradierten Gewohnheiten getrennter Verwaltung von Hardware-Einheiten bricht. Haben sich Lösungen wie „Sand“ oder „Sybase IQ“ noch auf spaltenbezogenes Speichern beschränkt, nutzen „Netezza“ oder „Teradata“ bereits Hardware-/Software-Pakete. Wie umkämpft dieses Segment ist, zeigen SAPs Kauf von Sybase, die Integration von DataAllegro in Microsofts SQL Server, der Aufkauf von Greenplum durch EMC oder jüngst der Kauf von Netezza durch IBM.

Doch die Exadata-Lösung geht mit großen Meilenstiefeln diesem Trend voran. Hardware und Software werden in einer einmaligen Art miteinander kombiniert. Es ist die Exadata-Storage-Software, die den Kick in der Hardware liefert. Auch hier spielt die Datenbank ihre genaue Kenntnis über die Speicherbedürfnisse von typischen Datenbank-Daten aus. Geschickt ist dabei, dass die klassische Oracle-OLTP+DWH-Datenbank mit all ihren mächtigen Funktionen beibehalten wird und die von Nischenanbietern zunächst als Alleinstellungsmerkmal entwickelten neuen Features jetzt auch in der Oracle-Datenbank enthalten sind:

- Die Verlagerung von Datenbank-Prozessen direkt in das Plattensystem (Exadata-Storage-Server). Damit ist Filtern direkt beim Plattenzugriff möglich, ohne dass alle Tabellendaten in die SGA geladen werden müssen.
- Ein neues Datentransfer-Protokoll zwischen Storage-Servern und Datenbank-Server, das ideal auf typische Tabellendaten abgestimmt ist. Die Datenbank verarbeitet die Daten direkt, ohne sie zwischenzupuffern und auch ohne die klassische Oracle-Block-Struktur zu nutzen.
- Ein neues, transparentes Storage-Index-Verfahren, das sich on-the-fly merkt, wo die am häufigsten genutzten Daten (Spalten und Werte-Bereiche) auf den Platten liegen.
- Ein neues, spaltenorientiertes Komprimierungsverfahren mit Kompressionsraten von 6 bis 40 minimiert IO-Zugriffe und steigert die Performance.

Alle zu Beginn beschriebenen Nachteile des separaten Storage-, Netz-, Server- und Datenbank-Managements werden durch Exadata aufgehoben. Das System besitzt sein eigenes „privates Netzwerk“ (40 GB schnelles Infiniband) und verfügt über massiv parallel eingesetzte und nur für die Datenbank reservierte 168 Platten (verwaltet durch ASM).

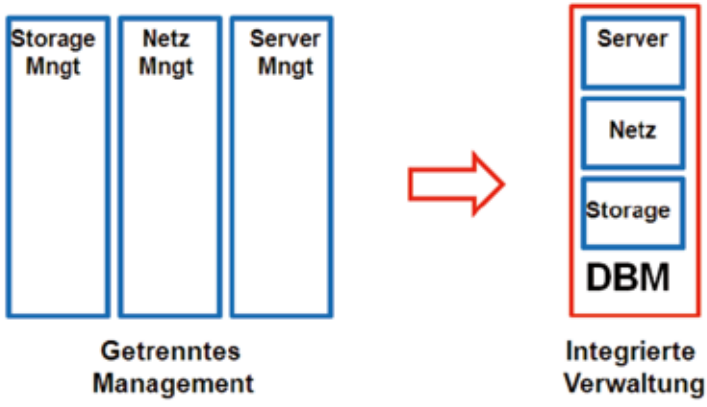

Abbildung 4: Die Vorteile von Exadata

Über Performance bei Exadata zu sprechen ist fast schon unfair. Denn bereits die kleinste Ausbaustufe (Quarter Rack) liefert mit 4,5 GB/sec Leseperformance ein Vielfaches dessen, was traditionelle Systeme schaffen (oft kleiner als 1GB/sec, kaum mehr als 2 GB/sec). Der Nutzen der extrem hohen Performance liegt dabei in dem Potenzial von fast unerschöpflichen Leistungsreserven, nicht in einer Verkürzung der Abfragezeit von vielleicht 8 auf 1 Sekunde. Die entscheidende Frage lautet nicht: „Wie schnell soll die Anwendung X laufen“, sondern „Wie viel zusätzliches Potenzial für mehr und weitere Aktivitäten ist gut für mein Unternehmen?“.

In einer Analyse fand einer der ersten Exadata-Anwender bezogen auf eine Testreihe von zehn Auswerte-Jobs heraus, dass diese auf ihrer alten Hardware 78,2 Stunden Laufzeit benötigten. Die gleichen Jobs waren auf der DBM aufaddiert in 7,8 Stunden fertig. Es blieben 70,4 Stunden freie Zeit. Was geschieht in dieser freien Rechenzeit? Es ist Kapazität für zusätzliche Dinge, an die zuvor nicht zu denken war.

Unterm Strich bewirkt die Database Machine in Unternehmen folgende Veränderungen:

- Anwender werden mutiger: Sie formulieren mehr und komplexere Abfragen
- Die IT-Abteilungen bekommen die Chance für mehr Optimierungen. Wo vorher zu 100 Prozent ausgelastete Systeme aus Angst vor einem Zusammenbruch nicht angefasst wurden, ist jetzt Raum für zusätzliche Tests und Optimierungsarbeit. Fehlläufe sind problemlos wiederholbar
- Die Entwicklung neuer IT-Services erfolgt schneller
- Die Online-Verfügbarkeit steigt
- Anwendungen und Laufzeiten werden berechenbarer (kalkulierbare Skalierung)
- Anwendungen werden ohne Umprogrammieren aus dem Stand schneller, wenn sie auf der Exadata laufen
- Der Betrieb mit all den komplexen Job-Netzen wird flexibler
- Durch die einheitliche Verwaltung von Storage, Netz und Server gestaltet sich der Betrieb schneller und einfacher

Die Themen und Lösungen zeigen, dass sich künftig auch in den IT-Abteilungen einiges im Umgang mit Storage und großen Datenmengen ändern wird. Die Entwicklung bleibt spannend.

Kontakt:
Alfred Schlaucher
alfred.schlaucher@oracle.com