Exadata V2 Erfahrungsbericht

Manfred Drozd
In&Out AG
Kilchbergsteig 13
CH-8038 Zürich

Schlüsselworte

Exadata V2, Architektur und Interna, Performance

Einleitung

Der Vortrag betrachtet schwerpunktmässig die Performance Eigenschaften der Exadata V2 Quarter Rack, wie sie im ersten Halbjahr 2010 an Kunden ausgeliefert wurde.

Exadata Database Server

Der Exadata Database Server ist ein Sun Fire X4170 Server, der auch im Internet in ähnlicher Konfiguration für weniger als 15‘000 USD bestellt werden kann. Er verfügt über zwei schnelle Quad-Core Xeon Prozessoren mit Multi-threading (cpu_count = 16). Die Hauptspeicherkapazität von 72 Gbyte (18 x 4 Gbyte DDR-3 DIMM) ist für wenige Datenbank Instanzen ausreichend. Kunden, welche die Exadata für die Konsolidierung einsetzen wollen, sollten die maximale Hauptspeicherkapazität von 144 Gbyte pro Database Server bestellen. Dies setzt den Einsatz der doppelt so teuren 8 Gbyte DIMMs voraus und wird wohl erst ab der nächsten Exadata Generation (ab September 2010) unterstützt. Zur Erinnerung: im x86 Umfeld kosten 128 Gbyte RAM (4 Gbyte DIMM) weniger als 10‘000 USD!

Der Database Server verwendet das Betriebssystem Oracle Enterprise Linux. Die Konfiguration des Database Servers entspricht weitgehend konventionellen Datenbank Servern. Es kommt Oracle 11.2.0.1.0 zum Einsatz. Daten sind auf dem Datenbank Server keine gespeichert. Speziell an diesem Server ist nur der Host-Channel-Adapter für InfiniBand, ansonsten unterscheidet sich dieser Server nicht von anderen konventionellen Datenbank Servern.

```
SQL> select * from v$cell;

CELL_PATH                      CELL_HASHVAL
------------------------------ ------------
192.168.10.3                    398250101
192.168.10.4                     88802347
192.168.10.5                   2520626383
```

Über die View v$cell erkennt der Datenbankserver die zur Verfügung stehenden Exadata Storage Cells und verwendet deren IP Adressen in v$asm_disk. Die Kommunikation zur Exadata Storage Cell erfolgt über das iDB Protokoll, das sowohl einfache I/O Funktionalität für blockweises Lesen und Schreiben umfasst, aber auch erweiterte Funktionen für Offload Funktionen und Ressource Management.

Mit folgenden init.ora Konfigurationsparametern kann das Verhalten des Exadata Database Server beeinflusst werden:

Parameter	Default Wert	Beschreibung
cell_partition_large_extents	TRUE	Ermöglicht *large extent allocation* für partitionierte Tabellen
cell_offload_processing	TRUE	Aktiviert SQL offload processing
cell_offload_decryption	TRUE	Aktiviert SQL offload encryption
cell_offload_compaction	ADAPTIVE	?
cell_offload_plan_display	AUTO	Zeigt beim EXPLAIN PLAN auch die offload Funktionen an

Da der Exadata Database Server eine Standard Komponente ist, erwarten wir hier für die Zukunft mehr Optionen bei der Konfiguration: andere Hauptspeicherkapazitäten, neuere und schnellere Prozessoren, Prozessoren mit 6 oder 8 Cores, eventuell sogar das Betriebssystem Solaris anstatt Oracle Enterprise Linux.

CPU Performance Tests

Obwohl die Exadata V2 nicht die neuesten und schnellsten Prozessoren hat, ist sie doch enorm leistungsfähig. Folgende Tabelle zeigt den Leistungsvergleich zu einem klassischen RSIC Server mit Power7 Prozessoren.

Anzahl Prozesse zur Last Erzeugung	**IBM POWER7** **1 Server** 2 cpus 3.5 GHz 12 cores, 48 threads Oracle Lizenz: 12 * 1.0 = 12 Prozessoren			**Intel E5540** **2 Server** 2 x 2 = 4 cpus 2.53 GHz 2 x 8 = 16 cores, 2 x 16 = 32 threads Oracle Lizenz: 16 * 0.5 = 8 Prozessoren		
	CPU Auslastung [%]	Durchsatz Total [kOps/sec] OraBench T132 Airthmetic Mix	Durchsatz pro Prozess [kOps/sec]	CPU Auslastung [%]	Durchsatz Total [kOps/sec] OraBench T132 Airthmetic Mix	Durchsatz pro Prozess [kOps/sec]
1	2.8	339	339	5.1	512	512
2	4.7	678	339	6.2	1‘000	500
4	8.3	1‘356	339	6.1	2‘000	500
8	15.0	2‘581	323	22.3	3‘809	476
16	25.6	3‘137	196	48.1	8‘421	526
32	55.4	4‘211	132	92.6	10‘491	327
64	87.2	4‘812	75	96.2	10‘847	169

Es fallen 3 Dinge auf:

- die multi-threaded Prozessoren zeigen die Auslastung nicht mehr korrekt an; dies gilt nicht nur für Intel Prozessoren, ähnliche Beobachtungen haben wir auch bei SPARC und POWER Prozessoren gemacht. Man vergleiche z.B. bei der Exadata den Leistungszuwachs zwischen Parallelität 16 und 32 und der damit verbundenen CPU Auslastung.
- Multi-threading bringt zwar noch einen Leistungszuwachs, hat im Oracle Umfeld aber nur begrenzte Bedeutung. Es wird erst in einem sehr hohen Auslastungsbereich wirksam (man vergleiche den Leistungszuwachs beider Prozessoren, wenn die Anzahl Prozesse die Anzahl Cores überschreitet) und geht zu Lasten des Durchsatzes bzw. der Servicezeit einzelner Prozesse (siehe Spalte „Durchsatz pro Prozess“).

- das Preis-/Leistungsverhältnis von x86 Prozessoren ist momentan unschlagbar. Sowohl vom Speed (Parallelität 1) als auch vom Durchsatz (Parallelität 32) ist der x86 Prozessor leistungsfähiger als der Power7 Prozessor. (Diese Aussage gilt übrigens auch im Vergleich von x86 Prozessoren zu Sparc und Itanium Prozessoren.) Wenn dann noch die Oracle Lizenzkosten und die Kosten des Servers ins Verhältnis zur Leistung gesetzt werden, schneidet der x86 Prozessor deutlich (Faktoren!) besser als alle anderen Prozessoren ab.

Exadata Storage Cell

Die Exadata Storage Cell ist ein Sun Fire X4275 Server (Listenpreis ca. 35‘000 USD). Dieser Server kann nicht mehr im Internet bestellt werden und wurde im August 2010 abgelöst. Der X4275 Server verfügt über die gleiche Prozessorleistung wie der Datenbankserver und hat eine Hauptspeicherkapazität von 24 Gbyte für den Cell Storage Index.

Das System kann in 2 Diskkonfigurationen geliefert werden:
- Kapazitätsoptimiert mit 12 SATA Laufwerken (7.2k rpm) mit je 2 Tbyte Kapazität
- Transaktionsoptimiert mit 12 SAS Laufwerken (15k rpm) mit je 600 Mbyte Kapazität

Auf der Exadata Storage Cell gibt es neben dem User root zwei weitere User: celladmin und cellmonitor (mit eingeschränkter shell), aber keinen User oracle.

Auf dem System laufen 3 Prozesse:
- CellServer (CELLSRV), multi-threaded Prozess, der für die eigentliche I/O Verarbeitung zuständig ist.
- Management Server (MS) für Konfiguration und Cell Management.
- Restart Server (RS) überwacht den Status der anderen beiden Prozesse.

Das Management einer Exadata Storage Cell erfolgt über ein separates Tools CELLCLI, das SQL*Plus ähnelt.

Wir haben bislang nur mit den transaktionsoptimierten Exadata Systemen gearbeitet. Bei unseren Performancetests konnten dabei ungewöhnliche Werte erreicht werden:
- Beim sequential write wurden 4 Gbyte/sec erreicht
- Beim sequential read wurden sogar 4.5 Gbyte/sec erreicht

Diese Werte wurden mit den konventionellen Platten erreicht, ohne Einwirkung des Flash Cache. Dies bedeutet eine durchschnittliche Lese- und Schreibleistung von über 100 Mbyte/sec pro Disk! Der Setup der Exadata Storage Cell ist also in der Lage, die Leistungsfähigkeit von Platten bis ans physikalische Limit zu nutzen. Bei konventionellen Plattformen mit ASM erreicht man typischerweise einen sequentiellen Durchsatz von 20 – 50 Mbyte pro Disk.

Auch die REDO Logfiles sind standardmässig auf konventionellen Platten abgelegt. REDO Logfiles auf Flash Technologie abzulegen bringt keine Vorteile, da REDO Logfiles sequentiell geschrieben werden und die Flash Technologie ja eher bei random I/O überlegen ist. Bei unseren Data Load Tests konnten je nach Test bis zu 30‘000 Commit pro Sekunde (*transactional data load*) bzw. bis zu 920 Gbyte pro Stunde (*bulk data load*) verarbeitet werden. Dies entspricht einer Laderate von ca. 250 Mbyte/sec.

Hardwaretechnisch zeichnet sich die Exadata Storage Cell durch einen Infiniband Host-Channel-Adapter und 4 Sun F20 SmartFlash Karten (je 96 Gbyte Kapazität) aus. Die SmartFlash Karten sind über PCI direkt am Memory Bus des Servers angeschlossen und garantieren so ein Höchstmass an Übertragungsleistung, die mit SAS oder FC Host Bus Adaptern nicht möglich wären. Die SmartFlash Karten können sowohl als superschnelle Disks oder als Write-Through Cache der 12 Disklaufwerke konfiguriert werden. Jede Exadata Storage Cell kann dank dieser Flash Technologie über 80‘000 IOPS (lesend) oder mehr als 4 Gbyte/sec (lesend) verarbeiten.

Die Anmerkung „Write-Through Cache“ ist nicht ganz unwichtig. Bei OLTP Systemen mit hoher Rate an ändernden Transaktionen wird der Durchsatz schnell durch die Anzahl Disks limitiert. Wir haben das bereits in unseren Benchmark nachweisen können. Eine Exadata QR verfügt nur über 36 Disks, d.h. mehr als 9‘000 IOPS können kaum erreicht werden.

Wenn die SmartFlash Karten als Cache genutzt werden, gibt es wiederum verschiedene Optionen:

- Man überlässt Oracle die Cache Verwaltung
- Man lädt gezielt Objekte in den Cache

```
SQL> create table test (a1 number);

Table created.

SQL> alter table test storage (cell_flash_cache keep);

Table altered.

SQL> alter table test storage (flash_cache keep);

Table altered.

SQL> alter table test cache;

Table altered.
```

Mit Oracle 11.2 und der Exadata stehen nun 3 verschiedene Möglichkeiten zur Verfügung, Speicherobjekte in unterschiedliche Arten von Caches zu pinnen (siehe Beispiel rechts). Das erste Statement bindet die Tabelle in den Cache der Exadata Storage Cell, das zweite Kommando bindet die Tabelle in den Flash Cache (neues Oracle 11.2 Feature, aber nur für Solaris und OEL) und das dritte Statement bindet die Tabelle in den Buffer Cache.

Für die unterschiedlichen Caches gelten auch stark unterschiedliche Servicezeiten:

- Zugriff im Buffer Cache der SGA im Nano Sekunden Bereich (10^{-9})
- Zugriff im Flash Cache oder im Cell Flash Cache im Mikro Sekunden Bereich (10^{-6})
- Zugriff auf konventionelle Platte im Milli Sekunden Bereich (10^{-3})

Applikationen, die eher I/O bound sind, werden sehr stark von der Flash Technologie profitieren. Applikationen, die eher CPU bound sind, werden kaum Verbesserungen gegenüber einem klassischen x86 Server wahrnehmen.

Bei unseren Benchmarks haben wir die Benchmarktabelle komplett in den Flash Cache laden können und anschliessend Performancezahlen gemessen, die das Attribut *extrem performance* wirklich verdienen:

- Beim sequential read konnte ein Spitzenwert von 12.5 Gbyte/sec erreicht werden
- Beim random read konnte ein Spitzenwert von 288‘000 IOPS erreicht werden.

Und das alles auf einer Exadata Einstiegskonfiguration (Quarter Rack) mit nur 3 Exadata Storage Cells!

Die Exdata Storage Cell wird wohl eher ein relativ geschlossenes System bleiben. In Zukunft können wir uns hier mehr Vielfalt insbesondere bei den Speichermedien vorstellen: unterschiedliche Kapzitäten und Geschwindigkeiten der Platten, verschiedene Storage Tiers.

Exadata Storage Cell Offload Funktionen

Während die Hardware Komponten auch für konventionelle Plattformen zur Verfügung stehen, kann die Exadata Storage Cell Software nicht separat gekauft werden. In dieser Software liegt die eigentliche Innovation:

- Extrem effiziente Nutzung der Hardware Ressourcen
- Optimierung der Datenbank Verarbeitung durch offload Funktionen

Auf die einzelnen Offload Funktionen gehen wir im zweiten Teil der Präsentation ein. Eine Übersicht, welche Offload Funktionen von der aktuelle Exadata Version unterstützt werden, findet man in der View V$SQLFN_METADATA (Spalte OFFLOADABLE).

Exadata Storage Area Network

Bei den I/O Performancezahlen wird schnell klar, dass konventionelle Storage Area Networks, die auf IP oder FC Technologie basieren, hoffnungslos überfordert sind oder zumindest einen massiven Einsatz von Hardware benötigen. Die Exadata verfügt quasi über ein eigenes internes Storage Area Network für eine extrem hohe Übertragungsleistung zwischen Datenbank- und Storage-Server. In einer Quarter-Rack Konfiguration kommen zwei InfiniBand Switches mit je 40 Gbit/sec Übertragungsleistung zum Einsatz. Die Latenzzeit beträgt 6 µsec (6×10^{-6} sec). InfiniBand wird auch für die Kommunikation zwischen RAC Knoten genutzt.

Extreme Performance Out-of-the-box

Die Exadata liefert ihre extreme Performance *out-of-the-box*. Wochenlanges Engineering, Testen und Optimieren der Plattform sind nicht mehr notwendig. Der Oracle Support benötigt ca. 2 – 3 Tage für die Installation der Exadata, anschliessend kann mit dem Einbinden der Exadata in die Datacenter Infrastruktur (Monitoring, Backup) und der Migration von Datenbanken begonnen werden. Dies ist zunächst eine schlechte Nachricht für Engineering Organisationen. Vermutlich wird sich deren Arbeitsschwerpunkt in Richtung Unterstützung der Anwendungsentwicklung verschieben. Die Vielfalt und Komplexität von Oracle Funktionen hat in den letzten Jahren rasant zugenommen. Die intelligente Nutzung dieser Features in der Applikationsentwicklung hinkt stark hinterher. Hier gibt es einen grossen Nachholbedarf, der in Zukunft von den Engineers bewältigt werden kann.

Hinweise für die Evaluation von Exadata Datenbank Maschinen

Für die Bewertung der Exadata Leistungsfähigkeit sind neue Kennzahlen notwendig. Wir vermissen in den Evaluationsprojekten häufig eine Bewertung folgender Aspekte:

Preis-/Leistungsverhältnis der Prozessoren:

- CPU Geschwindigkeit
- CPU Durchsatz
- Oracle Lizenzkosten
- Kosten für Ausbau des Hauptspeichers

Preis-/Leistungsverhältnis der Speicher:

- I/O Servicezeiten (z.B. für Call Center Anwendungen)
- Gbyte/USD
- IOPS/USD
- Mbps/USD
- IOPS pro Gbyte Speichermenge

Preis-/Leistungsverhältnis der Storage Area Network Infrastruktur

- Was würde es kosten, um das bestehende SAN so zu erweitern, um mit konventionellen Plattformen ähnliche Datenmengen bewegen zu können (Anzahl Host-Bus-Adapter, Kapazität Switches).

Performance Monitoring auf der Exadata

Die Exadata kann in das Monitoring des Enterprise Managers eingebunden werden. Auf der Exadata Storage Cell stehen via CellCLI weitere Kommandos zur Verfügung. Wir bevorzugen das Monitoring aus dem Datenbank Server via Data Dictionary Views.

Neue Statistiken in V$SYSSTAT

Um die Effizienz der Exadata Storage Cell zu überprüfen, wurden 39 neue Statistikwerte eingeführt. Hier einige wichtige Statistikwerte, die teilweise auch in den Views VSQL, VSQLAREA und V$SQLSTATS auftauchen:

Statistik Name	**Bedeutung**
`cell flash cache read hits`	Anzahl Leseoperationen, die im Flash Cache verarbeitet wurden (*cache hits*).
`cell physical IO bytes saved by storage index`	Anzahl Byte, die wegen dem Storage Index nicht gelesen werden mussten (*storage index block skipping*).
`cell physical IO bytes eligible for predicate offload`	Datenmenge in Byte, die für *predicate offload* Verarbeitung geeignet ist. Spalte IO_CELL_OFFLOAD_ELIGIBLE_BYTES in V$SQL.
`cell physical IO interconnect bytes returned by smart scan`	Anzahl Byte, die von *smart scans* zurückgeliefert wurden. Spalte IO_CELL_OFFLOAD_RETURNED_BYTES in V$SQL.
`cell physical IO interconnect bytes`	Anzahl Bytes, die zwischen Database Server und Storage Cell ausgetauscht wurden. Spalte IO_INTERCONNECT_BYTES in V$SQL.

Neue Events in V$SYSTEM_WAIT

Zur Performance Überwachung der I/O Aktivitäten auf der Storage Cell wurden ebenfalls einige neue *wait events* eingeführt.

Event Name	**Bedeutung**
`cell smart table scan`	Der Datenbank Server wartet auf den Abschluss eines *cell full table scan*
`cell smart index scan`	Der Datenbank Server wartet auf den Abschluss eines *cell fast full index scan* (Index oder IOT Segment)

`cell single block physical read`	Wie `db file sequential read` bei konventionellem Storage
`cell multiblock physical read`	Wie `db file scattered read` bei konventionellem Storage

Neue Hinweise in den Execution Plans

Auch die Beschreibung in den Execution Plans wurde erweitert. Operationen der Exadata Storage Cell werden separat ausgewiesen (in unserem Beispiel TABLE ACCESS STORAGE FULL):

```
-----------------------------------------------------
| Id  | Operation                    | Name         |
-----------------------------------------------------
|   0 | SELECT STATEMENT             |              |
|   1 |  SORT AGGREGATE              |              |
|   2 |   PX COORDINATOR             |              |
|   3 |    PX SEND QC (RANDOM)       | :TQ10000     |
|   4 |     SORT AGGREGATE           |              |
|   5 |      PX BLOCK ITERATOR       |              |
|   6 |       TABLE ACCESS STORAGE FULL| TAB_HEAP_RLP |
-----------------------------------------------------
```

Offload Funktion Incremental Backup

Für die offload Funktion *incremental backup* wurde die View V$BACKUP_DATAFILE um die Spalte BLOCKS_SKIPPED_IN_CELL ergänzt. Die Spalte zeigt an, wie viele Blöcke von der Storage Cell für ein i*ncremental backup* herausgefiltert wurden.

Komprimierungsverfahren

Mit unseren Benchmark Tabelle (1 Milliarde Datensätze, Zahlenwerte und Zeichenketten werden zufällig generiert) haben wir die verschiedenen Komprimierungsverfahren, die nun mit der Exadata zur Verfügung stehen, miteinander verglichen (siehe Tabelle rechts).

Komprimierungsverfahren	Kapazität der Tabelle
Ohne	240 Gbyte
BASIC (wie in Oracle 10)	56 Gbyte
QUERY HIGH (HCC)	17 Gbyte

Die Effizienz der Komprimierungsverfahren hat uns überrascht. Allerdings ist die CPU Auslastung nicht zu unterschätzen. Wir konnten bei unseren Tests die Prozessoren vollständig auslasten. Die ARCHIVE HIGH Option nimmt zudem auch eine überproportional lange Verarbeitungszeit in Anspruch. HCC wird übrigens auf den Datenbank Servern und nicht auf den Storage Cells durchgeführt.

Danksagungen

Ich möchte besonders Frau Karin Rüegg, CEO der Tradeware AG, für die hervorragenden Arbeitsmöglichkeiten im Schweizer Exadata Competence Center und Herrn Janos Horvath, IT Solution Architect bei Sun/Oracle, für die vielen fruchtbaren Diskussionen danken!

Literatur

[1] Oracle Exadata Storage Server Users Guide 11g Release 2; E13861-04, December 2009
[2] Oracle Exadata Storage Server Software Release Notes 11g Release 2, E13862-03, November 2009

Kontaktadresse:

Manfred Drozd
In&Out AG
Kilchbergsteig 13
CH-8038 Zürich

Telefon: +41 44 485 60 60
Fax: +41 44 485 60 68
E-Mail info@inout.ch
Internet: www.inout.ch, www.orabench.com, www.exadata.ch