Oracle Exadata Database Machine performance getest

Responstijd van 1 seconde

Herman Slange en Frits Hoogland

Oracle levert de Oracle Exadata Database Machine, een vernieuwende manier om hardware en software te combineren en performance, schaalbaarheid en beschikbaarheid naar een veel hoger niveau te tillen in vergelijking tot een configuratie die is opgebouwd uit losse componenten.

Deze unieke combinatie is interessant voor alle database applicatietoepassingen, of het nu gaat om Online Transaction Processing (OLTP) toepassingen, datawarehouse-toepassingen of een combinatie van die twee. Tevens is het concept van de databasemachine zeer interessant om kosten te besparen en beheer te vereenvoudigen door consolidatie.

Grid in a box

De Oracle Exadata Database Machine wordt ook wel 'Exadata' genoemd. Dit is ook de term die Oracle soms in de marketing gebruikt. De 'X' op de voorkant van het rack komt ook van 'Exadata'. Het gaat hier om een eenvoudig te deployen, schaalbare 'out-of-the-box' machine, geoptimaliseerd voor gebruik door de Oracle database software. De databasemachine is het best te omschrijven als een 'grid in a box', omdat hij bestaat uit een combinatie van database servers, storage servers een InfiniBand netwerkinfrastructuur voor onderlinge communicatie en andere componenten om het geheel te laten werken (zoals een KVM-console en een switch voor beheertoegang tot alle componenten en Lights-Out adapters).
De Oracle Exadata databasemachine is opgebouwd uit industriestandaard hardwarecomponenten van Sun en Oracle en deze componenten zijn op elkaar afgestemd met betrekking tot performance, beschikbaarheid en beheersbaarheid. Alle componenten zijn redundant uitgevoerd: beschikbaarheid is daarmee één van de ontwerpfundamenten. Afbeelding 1 toont hoe zo'n Oracle Exadata databasemachine er van binnen uitziet.

De bovenste laag, ook wel de 'upper half' genoemd, bevat database servers op basis van Intel architectuur met als besturingssysteem Oracle Enterprise Linux. De database- en clustersoftware is de standaard database- en grid infrastructuursoftware van Oracle versie 11.2, exact dezelfde versie als de van de Oracle website te downloaden versie. Dit maakt het eenvoudig en kostenbesparend om één of meerdere bestaande Oracle databases te migreren naar een Exadata databasemachine en zo direct te profiteren van alle voordelen van Exadata zonder wijzigingen te maken in de databaseconfiguratie.
De onderste laag, ook wel de 'lower half'genoemd, bevat storage servers. Dit zijn eveneens servers op basis van Intel architectuur met als besturingssysteem Oracle Enterprise Linux. Deze bevatten echter de Exadata Storage Server software en fungeren als dedicated storage voor de database servers in de 'upper half'. De Storage Server software kan, samen met optimalisaties in de Oracle database storagelaag (ASM), leiden tot grote verbeteringen in performance met betrekking tot niet alleen reguliere Oracle database implementaties, maar ook andere databases (zowel specialistische datawarehouse databases als reguliere databases).
De bovenste en de onderste laag zijn redundant gekoppeld via InfiniBand met twee InfiniBand switches. InfiniBand is een moderne netwerktopologie die met name voor I/O en clustering (high speed interconnect) gebruikt wordt. Er is voor InfiniBand gekozen omdat dit een veel grotere bandbreedte en veel lagere latency heeft dan Fibre Channel. Ter illustratie: een reguliere Fibre Channel adapter (HBA) heeft een throughput van 4 Gbps (Gigabit per seconde), terwijl de InfiniBand connecties in een databasemachine een throughput hebben van 40 Gbps.

Eenvoudige implementatie

Omdat de databasemachine volledig geassembleerd en geconfigureerd wordt geleverd, is implementatie in een datacenter snel te realiseren, waardoor de kosten en tijd voor het deployen van een databases tot een minimum beperkt worden. Dit lijkt triviaal, maar is een groot voordeel. Om een reguliere configuratie van losse servers, storage en netwerk voor optimale database performance te configureren, kan er aanzienlijke tijd worden gespendeerd aan het aansluiten van deze verschillende hardwarecomponenten. In veel organisaties wordt dit door verschillende teams uitgevoerd zoals hardware technici om de server te plaatsen, netwerk technici om het netwerk aan te sluiten en storage technici om de servers aan te sluiten op het SAN. Op dit punt zijn alleen nog maar de draadjes aan elkaar gekoppeld. Daarna moeten de besturingssystemen geïnstalleerd worden, waarbij soms additionele drivers of firmware van de hardware opgezocht en geïnstalleerd moeten worden. Als dat gerealiseerd

is, kan de database geïnstalleerd worden. Dit vereist wederom interactie met de beheerder van het besturingssysteem, en/of aanpassingen. Een Exadata databasemachine bevat alle hardware- en softwarecomponenten en alles is op elkaar afgestemd.

Kostenbesparing

De Exadata databasemachine bevat optimalisaties voor zowel OLTP, DWH of een mix van beide workloads. Het feit dat op een enkele Oracle Exadata databasemachine meerdere databases geconsolideerd kunnen worden, maakt het voor een datacenter een interessant concept. Beheer en onderhoud van meerdere databases worden hierdoor eenvoudiger, overzichtelijker en kostenefficiënt. Door de op elkaar afgestemde, moderne componenten is het mogelijk dat Exadata een reeks van bestaande servers vervangt en daardoor is fysiek minder ruimte, minder stroomverbruik en minder koeling nodig in een datacenter.

Weinig tijd, veel data I/O verwerken

De drijvende kracht achter de performance van de databasemachine is de Exadata Storage Server. De Exadata Storage Server beschikt over specifieke mogelijkheden om veel data I/O intelligent te kunnen verwerken:

- Flash memory in de Exadata Storage Server kan gebruikt worden als cache om fysieke I/O te voorkomen. Oracle gebruikt bewust flash memory en geen flash disks, om daarmee een disk controller te elimineren als potentiële bottleneck;
- Exadata Smart Scans om dataprocessing efficiënt te offloaden naar de storage servers. Offloaden betekent dat een gedeelte van de processing van de database wordt uitgevoerd door de Exadata storage, en een resultaat wordt teruggegeven aan de database, in plaats van de diskinformatie;
- Storage indexes om I/O te elimineren op de storage servers. Een storage index wordt automatisch, zonder interactie met de databaselaag en dus ook zonder beheerders opgebouwd en automatisch toegepast. Toepassen betekent dat de minimum- en maximumwaarden van de velden in een blok van 1 MB aan databaseblokken worden bijgehouden, en niet worden gelezen als de predicates in de query dit uitsluiten. De I/O resource manager wordt gebruikt om I/O te kunnen sturen tussen verschillende databases of tussen services in de database.

Enorme opslagcapaciteit

Datacompressie kan leiden tot een aanzienlijke verlaging van de gebruikte opslagcapaciteit. De Oracle database kent al compressie op basis van het elimineren van dubbele waarden in een databaseblok. Een Exadata databasemachine heeft de mogelijkheid om Exadata-specifieke compressie toe te passen, genaamd Exadata Hybrid Columnar Compression of EHCC. Door het gebruik van EHCC kan een hoger niveau van datacompressie behaald worden dan met reguliere Oracle database compressietechnieken. Hierdoor kan een nog grotere besparing van opslagcapaciteit gerealiseerd worden. Het gebruik van EHCC (maar ook compressie in het algemeen) reduceert I/O, wat tot verbetering van performance kan leiden en kostenverlagend is. In een omgeving waarbij de I/O hoeveelheid en de performance zeer belangrijk zijn, kunnen de I/O resources specifiek verdeeld worden binnen de database en tussen de databases, zodat performance voorspelbaar en beheersbaar wordt. Dit wordt gerealiseerd door de database resource manager en de Exadata IO resource manager.

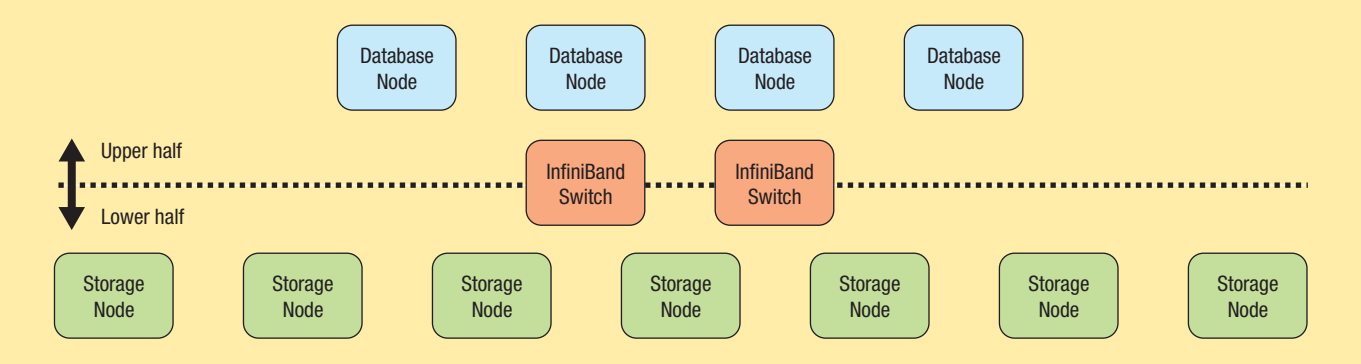
Van theorie naar praktijk

Dit verhaal klinkt mooi, maar waar zit nu precies het verschil? Of is dit de typische marketing over 'the latest and greatest'? Het verschil dat Exadata maakt zit in de mogelijkheid om bepaalde onderdelen van database query's te kunnen laten uitvoeren ('offloaden') door de Exadata Storage Servers. Dit betekent dat Oracle met Exadata dus database en storage gedeeltelijk geïntegreerd heeft. Als voorbeeld:

```
'select count(*) from large_table'.
```

Dit is een gesimplificeerde beschrijving van hoe een 'reguliere' database een tabel scant:

1. Segment header van 'large_table' wordt opgehaald, waar in staat welke databaseblokken zijn gealloceerd door 'large_table';
2. Een of meerdere blokken worden van disk gelezen en de inhoud wordt gelezen (en de rijen geteld in dit geval, vanwege 'count(*)');
3. Als er nog meer blokken zijn, wordt punt 2 herhaald totdat alle blokken zijn gelezen.



Afbeelding 1: Overzicht Oracle Exadata Database Machine.

Voor iedere keer dat een blok (of meerdere blokken, een 'multi-block I/O') opgehaald moet worden van disk wordt de doorlooptijd vergroot met de disk I/O latency. Dit is een gesimplificeerde beschrijving van hoe een Exadata 'Smart Scan' een tabel scant:

1. Segment header van 'large_table' wordt opgehaald, waar in staat welke databaseblokken zijn gealloceerd door 'large_table'. Als er veel blokken in de database buffercache staan, kan besloten worden de blokken regulier (zie boven) te scannen, omdat dat sneller is;
2. Er wordt besloten tot een 'Smart Scan'. Alle storage servers worden geïdentificeerd aan de hand van de blokinformatie;
3. Er wordt een ontvangst- en verzendkanaal opgezet naar iedere betrokken storage server;
4. De requests worden naar de storage servers verzonden door het database server process;
5. Het database serverproces zorgt ervoor dat er steeds iets meer te verwerken resultaten van de requests worden verstuurd door de storage servers, zodat het database serverproces de antwoorden continu kan verwerken. Hierdoor ondervindt een smartscan zo goed als geen vertraging van disk I/O latency.

Van 11 minuten naar 1 seconde

Dit is nog steeds de theorie. Laten we nu eens kijken wat dat 'in het echt' betekent. VX Company heeft in een democentrum een Exadata databasemachine ingericht zodat bedrijven met hun eigen dataset de kracht van Exadata kunnen uitproberen. In een Proof-of-Concept wordt precies duidelijk wat de voordelen van Exadata zijn in een specifieke klantsituatie. In het volgende voorbeeld hebben we de Exadata databasemachine van VX Company gebruikt voor een aantal testscenario's. Laten we beide mechanismen eens toepassen en naar de verschillen kijken aan de hand van een tabel met: grootte 133.425.004.544 bytes; 2.228 extents; 16.287.232 blokken.

1. Tabel lezen op de reguliere manier:
   Totale doorlooptijd: 695 seconden (ruim 11 minuten);
   Waarvan CPU tijd: 45 seconden;
   Waarvan I/O tijd: 652 seconden;
   (aantal I/O's: 127.238).
   Deze actie is duidelijk gelimiteerd door I/O. Is dit een goede tijd? Ja, dit is een goede tijd voor een dergelijke query. De I/O tijd is 652 seconden, als we dit delen door het aantal I/O's (127.238) dan krijgen we de gemiddelde I/O tijd per I/O: 652/127238 = 0,005 seconden. Dat is gemiddeld 5 milliseconden per I/O. Dat is een zeer goede tijd (de brochure van de Seagate Cheetah 15K.7 disk, die in een high performance Exadata V2 configuratie zit, vermeldt een gemiddelde leestijd van 5,4 milliseconden). Probeert u eens een I/O tijdcalculatie op uw huidige systeem te doen.
2. Tabel lezen via Exadata 'Smart Scan':
   Totale doorlooptijd: 40 seconden;
   Waarvan CPU tijd: 34 seconden;
   Waarvan I/O tijd: 6 seconden.
   Het verschil ten opzichte van de reguliere scan is duidelijk: 695 seconden versus 40 seconden via 'Smart Scan'. Ook is goed te zien dat nu de CPU de bottleneck is geworden.

Om op een Oracle Exadata databasemachine de 'Smart Scan' te kunnen gebruiken is *geen enkele* handeling nodig: als een database naar Exadata is gemigreerd, vindt de Smart Scan automatisch plaats. EHCC, Exadata Hybrid Columnar Compression moet wel specifiek geconfigureerd worden. De meest gebruikte strategie is om tabel (en/of index) partities te laten 'afkoelen' (wat betekent dat de data niet meer gewijzigd worden, of 'bewegen') en daarna EHCC op de partitie toe te passen. Als data statisch zijn, kunnen ze uiteraard meteen toegepast worden. De beperkingen van EHCC zijn dezelfde als de andere vormen van compressie in de Oracle database.
Laten we dit ook eens 'in het echt' bekijken. In dit geval maken we gebruik van de tabel die ook gebruikt is bij punten 1 en 2 hierboven, maar is EHCC 'Query compression' gebruikt. Dit is compressie die geoptimaliseerd is voor query performance. Naast 'Query compression' kent Exadata ook 'Archive compression', welke geoptimaliseerd is voor compressie, dus reductie van de grootte van het gecomprimeerde object.
De flash memory in de Storage Server is daarnaast geconfigureerd als disk, en de EHCC tabel is daarop geplaatst. Dit voorbeeld is uitgevoerd op de half rack Oracle Exadata databasemachine bij VX Company. Deze configuratie bestaat uit: vier database servers, met ieder twee Quad-Core Intel Nehalem CPU's; zeven storage servers, met ieder twee Quad-Core Intel Nehalem CPU's.

Om zo veel mogelijk van de aanwezige CPU's te gebruiken, wordt daarnaast gebruik gemaakt van de Parallel Query option, die er voor zorgt dat de scan voor de tabel wordt opgedeeld in ranges en elke range door een eigen Parallel Query 'slave' proces afgehandeld. Smart scans kunnen ook gedaan worden door Parallel Query Slaves.
Door EHCC Query mode wordt de grootte van de tabel van 133 GB gereduceerd tot 11 GB. Een reductie van maar liefst 92 procent! In dit geval is er gebruik gemaakt van 64 Parallel Query Slaves, verdeeld over de vier database servers.
Totale gebruikte tijd van alle processen: 12 seconden.
Totale CPU tijd van alle processen: 10 seconden.
Totale I/O tijd van alle processen: 2 seconden.

De responstijd van deze query waarbij gebruik gemaakt is van bovenstaande standaard technieken die Exadata biedt, is: 1 seconde! Dat is spectaculair te noemen: van 11 minuten naar 1 seconde door slim gebruik te maken van de beschikbare features. Nogmaals: de query is in alle gevallen niet veranderd.

Herman Slange (hslange@vxcompany.com) is Technisch Manager bij VX Company. **Frits Hoogland** (fhoogland@vxcompany.com) is Principal Consultant bij VX Company.