Exadata Database Machine - Die Konsolidierungsplattform

Malthe Griesel
Paragon Data GmbH
Friedrichsdorf

Schlüsselworte
Oracle Exadata Databasemaschine, Platinum Support, Hochverfügbarkeit, Engineered Systems

Einleitung

Paragon Data betreibt für den Kunden DBH und andere am Standort Friedrichsdorf ein BSI zertifiziertes Rechenzentrum. Die Paragon betreibt und betreut für den Kunden DBH unter anderen auch mehrere Datenbanken (Warenwirtschaft, Datawarehouse/ Managementinformationssystem, Kundenkontaktmanagement, Katalog und viele kleine Datenbanken).
Für diese Applikationen bestehen maximale Verfügbarkeiten zu den Ladenöffnungszeiten der Filialen, d. h. Montags bis Samstags von 08:00 Uhr bis 22:00 Uhr, maximale Höchstverfügbarkeit. Steht das System, kann kein Geld verdient werden. Zu alledem wird auf ein sehr gutes Antwortzeitverhalten der Applikationen sehr viel Wert gelegt.
Im Grunde genommen wird auf ein Oracle MAA-konformes System gesetzt, allerdings soll noch alles bezahlbar sein und von einem kleinen DBA-/ Linux-/ Storage und Netzwerkteam betrieben werden.
In der Vergangenheit hatte Paragon als einer der ersten einen POC mit Oracle und einer Exadta V1 in Readding durchgeführt. Die Ergebnisse waren durchweg positiv, sogar überragend. Leider scheiterte die Anschaffung einer Exadata für Paragon Data an kaufmännischen Gründen. Die Exadata war damals vor dem Hintergrund massiver Investitionen in ein Netapp 6080 Metrocluster einfach nicht sinnvoll.
In diesem Zuge wurde ein 8 Knoten RAC mit sehr leistungsfähigen Knoten (4 Cores pro Knoten, 192 GB Ram, 2 x 8 Gbit FC-Anbindung, Infiniband Interconnect) aufgebaut. Als Betriebssystem kam Oracle Enterprise Linux 5.7 mit dem UEK-Kernel zum Einsatz. Geplant war es, dieses sehr stark an die Exadata Databasemaschine angelehnte System mit damals noch optional zu erwerbenden Storageservern auszustatten. Mit der Übernahme Oracles von Sun und dem dadurch bedingten Systemwechsel von HP zu Sun/ Oracle war dieses Vorhaben nicht mehr möglich.
Im Rahmen eines Projektes zur Erneuerung des Storagesystems gut 4 Jahre später kamen neben den typischen Mitbewerbern wie Netapp, Fujitsu und EMC auch Oracle Exadata mit ins Spiel, da ein Großteil des Storage für Datenbanken benutzt wird und bei Einsatz von Exadata für die Datenbanken noch mehrere kleinere vorhandene Netappsysteme weiter betrieben werden konnten.
Beim Betrieb einer Exadata versprachen wir uns außerdem personelle Ressourcen einsparen zu können, alleine weil die Anzahl der Server weniger wird und im Rahmen des Platinum Supports z. B. das Patching für uns sehr viel einfacher wird, da OS, Clusterware, Firmware der Switches und die wichtigsten DB's durch Oracle Spezialisten im Rahmen des Platinum Supports patchen zu lassen.
Schließlich und letztlich war das Exadataangebot in dem gegebenen Kontext unter Berücksichtigung unserer bestehenden Netapp Landschaft das wirtschaftlichste.

Die alte Systemlandschaft

Wie schon einleitend erwähnt betrieb die Paragon Data für die DBH mehre Datenbanksubsysteme, darunter eine acht Knoten RAC mit einem Infiniband Interconnect und sehr gut ausgestattet, das bedeutet im einzelnen zwei Dualcore-Prozesoren, 196 GB RAM und 4 4GBit Fibrechannel Anbindung zur Netapp. Als Interconnect kam ein Infinband Netzwerk mit RDS als Protokoll zum Einsatz.
Mit diesem acht Knoten RAC wurde schon das selbe Ziel verfolgt, Hardware, Lizenzen bzw. Lizenzoptionen (Enterprise, RAC, Partitioning) zu konsolidieren. Leider ließen sich aus Ressourcenenpässen nicht alle Datenbanksysteme des Kunden DBH auf dieses System konsolidieren. Besondere Engpässe entstanden gerade im Umfeld von IO gegen die Netapp. Hier wurde sehr viel mit Hilfe des Netapp Professional Service versucht, allerdings schafften wir es nicht die sog. single

threaded Performance, die Performance eines einzelnen lesenden Prozesses, nicht mehr als auf 20 MB/s zu steigern. Dieses ließ sich zwar in Datenbanken wie dem Datawarehouse durch geeigneten Einsatz des Parallel Query Option steigern, allerdings die mögliche Leistung unter Benutzung beider Clusterknoten des Netapp 6080 Clusters von 1 GB/s haben wir nie unter Produktionsbedingen mit einer Datenbank erreicht. Gründe hierfür sind, dass die Netapp nicht ausschließlich Datenbanken bediente sondern auch noch zusätzlich in nicht unerheblichem Maß Filesharing Dienste wie SMB und NFS bediente. Im Endausbau lief das Netapp Cluster mit mehr als 500 Fibrechannel Festplatten.
Große Probleme bereitete die Netapp den DBA beim Durchführen von Backups. Die Backupstrategie sah in der Woche von Montags bis Samstag inkrementelle und Sonntags vollständige Backup mittels des Oracle RMAN auf ein Netapp NFS-Device vor. Das erstellen der Vollsicherungen stellte die Verantwortlichen vor Probleme, da die Sicherung mit mehr als acht Stunden das dafür vorgesehene Zeitfenster deutlich überschritt. Hierbei muss angemerkt werden, dass die Datenbanken der DBH (z. B. Warenwirtschaft und Datawarehouse) sehr stark wachsen (teilweise mehr als 30% Wachstum).
Hier hat das von Netapp propagierte Tool Snapmanager für Oracle (SMO) für das schnelle Backup nur bedingt Abhilfe geschaffen. Der SMO kann nur in sehr eingeschränkten Umfang mit ASM als Filesystem für RAC Datenbanken umgehen, NFS als Filesystem ist kein Problem und funktioniert mit SMO hervorragend. Die damals eingesetzte Version von SMO mit ASM ist in der Lage unter Last ASM und somit das Cluster zum Absturz zu bringen.
Während des gesamten Betriebes des 8 Knoten Clusters über mehr als vier Jahre kämpften die Linux- und Datenbankadministratoren mit massiven Stabilitätsproblemen. Das Cluster startete selbstständig immer wieder in unregelmäßigen Abständen mindestens einen oder mehrere Knoten durch. Sehr viele und teilweise stark eskalierte Servicerequests haben leider keine Lösung gebracht. Durch Updates des Kernels und der Clusterware sind zwar partielle Verbesserungen erreicht worden, aber richtig glücklich waren wir immer noch nicht. Der Fehler war immer im Umfeld von Infiniband oder Netapp und Multipathing des Betriebssystems zu suchen. Die Zusammenarbeit mit dem Oracle Support gestaltete sich als schwierig, da bei der Fehlersuche häufig Fingerpointing in Richtung Netapp oder Betriebssystem betrieben wurde. Letzteres war total unverständlich, da wir Oracle Enterprise Linux als OS eingesetzt und auch den passenden Support beauftragt hatten. Ein Zusammenbringen der Supportmitarbeiter von Netapp und Oracle war in nur ganz wenigen Fällen möglich, dann allerdings auch von Erfolg gekrönt. Allerdings wurde auch von Netapp auch gerne auf Fehlfunktionen des QLogic San Switches verwiesen. Dieses Support Chaos sollte durchbrochen werden.

Der Erneuerungsprozess

Im Zuge eines neu anzuschaffenden Storage Systems wurde die Oracle Exadata Database Machine durch die DBA zu den bekannten Storage Herstellern Netapp, EMC und Fujitsu mit ins Spiel gebracht. Nach einem sehr professionellen Sales und Presales Termin der Firma Oracle in den Räumen der Paragon Data in Friedrichsdorf begann der Verhandlungsprozess mit den einzelnen Storage Herstellern und Oracle.
Im Zuge einer Exadata Einführung war es möglich vorhandene Netapps (6 FAS 3240) für die Filesharing- und restlichen Aufgaben (SMB, NFS, VMWare, ...) zu nutzen. Für die Exadata sprachen außerdem, dass wir durch den Einsatz eines Exadata Quarterracks Zeit der DBAs einsparen würden, da zum einen hoffentlich nicht mehr mit so vielen unplaned Downtimes und deren Behebung wie mit dem acht Knoten RAC zu rechnen ist. Außerdem sollte die Exadata um ein Vielfaches schneller sein, sodass ausgiebige Tuningprozesse für spezielle SQL-Statements nicht mehr notwendig sein sollten. Diese Erkenntnis stammt aus dem POC vier Jahre zuvor. Als weiterer Pluspunkt für eine Exadata wurde der bessere Support für Engineered Systems mit dem Support für alles (OS, DB, Clusterware, Hardware) und besonders die versprochenen Leistungen des Platinum Supports für Engineered Systems mit vierteljährlichen Patchen und 7 x 24h Monitoring angerechnet. Hier lagen schließlich in der Vergangenheit arge Probleme, die für uns ggf. unlösbar waren.

Nach diesem Entscheidungs- und Bewertungsprozess blieben nach rund 2 Monaten intensiver Arbeit nur noch EMC, Netapp und Oracle im Rennen. Schließlich und letztlich haben sich die Gesellschafter der DBH nach unser Empfehlung für ein Oracle Exadata Quarter Rack entschieden.

Vorbereitungen zur Anlieferung der Exadata

Nach der Entscheidung der Gesellschafter der DBH für eine Exadata und der damit einhergehenden Bestellung wurde das System rund 6 Wochen später geliefert. Die Lieferung erfolgte nach einem Marathon an Gesprächen rund um die Auslieferung. Einige Telefonkonferenzen und ein sog Site Survey mussten durchgeführt werden, bis es schließlich zu einer Auslieferung der Exadata im Februar kam. Sehr viele Diskussionen wurden um die Themen „Reracking“ und Platinum Support und das dazugehörige Gateway geführt.

Ein Reracking war in unserem Rechenzentrum notwendig, da eine Tür mit nur 1,96m Höhe zu Niedrig für das 2m hohe System war. Auch wurden Rampen vermessen, die dann 0,3° zu steil waren. Alles wurde im schummrigen Licht mit einem Zollstock vermessen, so dass mit Messfehlern zu Rechnen war. Die Diskussionen rund um diese nicht Exadata konformen Wege in unser BSI-zertifiziertes Rechenzentrum waren sehr intensiv und nervenaufreibend. Allerdings nach rund 6 Stunden nach der Anlieferung stand die Exadata in unserem Rechenzentrum.

Das Gateway für den Platinum Support hat mindestens genauso viel Diskussionen verursacht. Nicht nur musste sich unsere Netzwerkabteilung mit den gewünschten Zugängen zum Gateway und von dort aus zum Exadatasystem kümmern und realisieren. Viel mehr Anstrengungen machte die Anforderung für den Rechner für das Gateway (64 Bit, 500 GB Plattenplatz, Quadcore, ...). Aus diesen Anforderungen schlossen wir, dass mit einem etwas älteren vorhandenen System das Gateway betreiben können würden. Die war jedoch ein Trugschluss, da die Installation durch Images geschieht und sehr neue Hardware voraussetzt. So mussten wir einen sehr neuen Server noch mit sehr teueren Festplatten ausstatten, um das Gateway aufbauen lassen zu können. Dieses System wurde innerhalb von Stunden bereitgestellt, die Platten mussten von einem Distributor 50 km entfernt abgeholt werden, um rechtzeitig mit der Installation beginnen zu können.

Migration von 11 Datenbanken in 6 Wochen

Neben der Organisation der Anlieferung und des Platinum Gateways haben die DBA die 6 Wochen Zeit zur Anlieferung sehr intensiv genutzt, die Migration von 12 Datenbanken zu Planen. Es wurden sehr genaue Pläne/ Checklisten für die Migrationen erarbeitet. Diese Listen beinhalteten u. a. die zu migrierenden Daten, zu notierende Laufzeiten, zu benutzende Skripte für Export, Import und Analyse. Weiterhin wurden die zu migrierenden Datenbanken sehr genau betrachtet, Parametereinstellungen, Größe der Tablespaces und deren Namen, Besonderheiten wie Oracle Text wurden aufgenommen und die Skripte für die Erstellung der Datenbanken wurden im Vorhinein erstellt.

Nach der Auslieferung wurde die Exadata hard- und softwareseitig innerhalb von rund 5 Tagen aufgebaut und an die DBA übergeben. Danach bauten die DBA das System auf, um 12 Datenbanken auf der Exadata zu betreiben.

Die Vorgehensweise hierzu war für alle Datenbanken die gleiche:

1. Datenbanksoftware mit einem eigenen OS-User installieren
2. Datenbank mit den bereits vorhandenen Skripten aufbauen und im Cluster integrieren
3. Backup- und Exportskripte einführen und testen
4. Datenübernahmeskripte (Export/ Import bzw. expdp/ impdp) testen und mehrfach ausführen
5. Fachabteilungen mit der neuen Datenbank auf der Exadata testen lassen
6. Finale Datenübernahme in einem dafür ausgemachten Zeitfenster durchfühen

Dieses stringente Verfahren wurde für alle Datenbanken gleich ausgeführt. Das Export-/ Importverfahren bietet sich im Exadataumfeld gerade an. Große Datenbanken wie unser Datawarehouse mit mehr als 5 TB Größe bis hin zur 4 GB großen Datenbank für das

Zeiterfassungssystem wurden in maximal 5 Stunden vom Altsystem auf die Exadata übernommen. Die Importe für die kleinen Datenbanken liefen nur wenige Minuten.

Anlaufschwierigkeiten

Nachdem Datenbank für Datenbank übernommen wurde, bekamen wir schon bekannte Probleme, einer der beiden Knoten startete unvorhergesehen in unregelmäßigen Zeitabständen durch.

Nachdem ein sehr hoch priorisierter Servicerequest sowohl durch Mitarbeiter von Oracle als auch durch Paragon Data eskaliert wurden, wurden als erstes eine Infinibandkarte aufgrund unnatürlicher Fehlermeldungen getauscht, diese Ämderung brachte nur bedingt Abhilfe. Stabilität wurde durch Umkonfiguration der Datenbanken, im Speziellen der Memoryparameter erreicht. Große Schwierigkeiten machte uns, dass wir die Memorykonfigurationen des Acht-Knoten-RAC aus der Vergangenheit übernommen hatten. Durch die acht Rechner mit rund 200 GB RAM konnten und mussten wir den Datenbanken sehr viel Memory zuweisen, um den schlechten IO unserer Netapp auszugleichen. Erst nachdem wir die SGA's der Datenbanken deutlich reduziert hatten Kernelparameter im Zusammenhang mit den mit der Exadata propagierten Hugepages angepasst hatten wurde die Stabilität des Systems gut. Hier sehe ich auch einen Nachteil der Exadata als Konsolidierungsplattform, soll eine Datenbank auf der Exadata hinzugefügt werden, muss die Konfiguration der Kernelparameter und der Hugepages angepasst werden (Doc ID 401749.1, Kernelparameter vm.nr_hugepages).

Nach dieser Änderung war die gewünschte Stabilität erreicht. Der Prozess für die Erreichung dieser Stabilität hat rund 14 Tage angedauert. Während dieser Zeit sank die Akzeptanz für die Exadata Database Machine im Unternehmen bei de Kollegen teilweise sehr dramatisch.

Ein weiterer Punkt der die Akzeptanz der Kollegen für die Exadata senken ließ, waren teilweise extreme Performanceprobleme in den Prozessen unseres Datawarehouses. Ein Prozess arbeitete füher mit allen acht Knoten parallel, nutzte den vorhandenen RAM und die Netapp Performance gut aus und lief rund 20 Minuten. Diesen Prozess haben wir ohne Änderungen übertragen auf die Exadata, danach lief dieser Prozess rund 48 Stunden. Im Zuge des Tuning dieses Prozesses haben wir sehr viel an den Parametern der Parallel Query Option versucht zu drehen, leider nur mit mäßigem Erfolg für das Ergebnis. Hier hat ein Servicerequest bei Oracle sehr wenig gebracht, obwohl sehr hoch priorisiert und durch mehrere Instanzen eskaliert (oracle und wir selber).

Abhilfe schaffte hier ein Workshop mit dem Partner Anykey und dem Mitarbeiter Stefan Agel. Das Statement konnte nicht die speziellen Eigenschaften der Exadata (Offloading) nutzen. Durch Umprogrammieren des Statements schafften wir es zusammen, das Statement von 48 Stunden auf 9 Sekunden zu beschleunigen.

Das Prinzip des Tuning lag darin, die Spezialdisziplin der Exadata in den Vordergrund treten zu lassen, lass die Datenbank soviel wie möglich auf Tabellen im Full-Table-Scan allerdings nur einmal lesen. Beim Lesen von Daten über einen Index kann die Exadata nicht ihre Performancevorteile gegenüber normalen Datenbanksystemen ausspielen.

Monitoring

Das Monitoring des Altsystems und der Exadata wurde bzw. wird mittels Oracle Grid Control 12c und Nagios durchgeführt. Die wichtigen Teile der Exadata wie Hardware, OS und Datenbanken (Tablespaces, Anazahl Sessions, ...) werden bei Paragon per Nagios überwacht. In Zukunft ist es geplant die Storageserver auch per Nagios zu Überwachen, leider gestaltet sich das sehr schwierig, da keine Software auf diesen Servern installiert werden darf, auch kein nrpe-Client für den Betrieb von Nagios.

Hier entwickeln wir gerade an einer Lösung, die mit den durch Oracle vorgegebenen Einschränkungen umgehen kann.

Das Performance Monitoring und Troubleshooting wir mittels Grid Control in bekannter Weise durchgeführt und führt sehr häufig sehr schnell zum Ziel. Es wird vom DBA Übung im Umgang mit dem Tool vorausgesetzt, um gute und schnelle Ergebnisse zu erzielen. Grid Control ist für den Betrieb von Exadata ein Muss, im Speziellen sind es die Management Packs Performance Monitoring und Changemanagement.
Das von Oracle propagierte Exadata Plugin für Gridcontrol liefert sehr schnell sehr ansehnliche gute Ergebnisse.

Fazit
Exadata stellt in meinen Augen mit seiner Skalierbarkeit eine gute Plattform für gerade mittelständische Kunden dar. Viele Mittelständler betreiben mehrere Datenbanken ggf. auf mehreren Systemen. Hier besteht beim Einsatz von Oracle Exadata Machines die Chance sehr gut zu konsolidieren, beispielsweise lassen sich vorhandene Datenbank Lizenzen schon benutzen, um eine Exadata zu betreiben. So war es in unserem Fall, 32 CPU waren im Cluster schon lizensiert, somit konnten wir den Posten Datenbanklizenzen beim Kauf des Viertel Racks außen vor lassen, es mussten „nur“ die Hardware und die Lizenzen für die Storageserver angeschafft werden. In unserem Fall waren das Hardware und 36 zu lizensierende Platten der Storageserver.
Die Themen Performanceprobleme sollten mit einer Exadata bei Ausnutzung der Exadataspezialitäten (Offloading) absolut kein Thema sein, gerade bei Eigenentwicklungen hat man hier sehr große Vorteile. Aber auch 3rd Party Software lässt sich für den performanten Betrieb mit Exadata tunen, hier ist Aufwand in der Untersuchung notwendig. Ein probates Mittel bei uns, war es geeignete Indizes zu löschen, um damit Prozesse dramatisch zu beschleunigen. Auf jeden Fall sollte eine Exadata genauso schnell sein, wie ein vergleichbares Datenbanksystem ohne die speziellen Exadata Storageserver.
Durch den hoffentlich erreichten Performncevorteil einer Exadata werden auch Ressourcen gespart, d. h. die DBA brauchen sich nicht mehr um das sehr schwierige Thema Performance Tuning kümmern. Eine Exadata, wenn denn die Prozesse etabliert sind, läuft ruhig vor sich hin, sofern keine technischen defekte vorliegen.
Mit dem Erwerb eines Enginered Systems wie eine Exadata haben Sie die Möglichkeit den Platinum Support zu nutzen. Dieser Support schließt zum einen das Monitoring und Verfolgen von Fehlern (Hard- und Software) via ASR (Automatic Service Request) und einem speziellen Platinum Support Gateway ein. Außerdem wird Ihr System nach Planung durch Sie durch einen Oracle Mitarbeiter in regelmäßigen Abständen gepatched. Hier wird viel Aufwand bei den DBA gespart, da Sie sich nicht immer in den Patchingprozess einarbeiten müssen, weil dieser durch einen geeigneten, geübten Oracle Mitarbeiter durchgeführt wird.
Generell gilt auch, dass Supportanfragen auf jeden Fall durch einen Oracle Mitarbeiter bearbeitet und gelöst werden. Sowohl Hardware als OS und Datenbanksoftware stammen aus dem Hause Oracle, bei schwierigen Servicerequests arbeiten unter Aufsicht immer gleich mehrere Teams bei der Lösung von Problemen zusammen, sodass in sehr schneller Zeit sehr gute Ergebnisse erzielt werden. Der allgemeine Support für Enginered Systems wie Exadata ist deutlich besser.
Für mich sind die Vorteile der Exadata, gerade hier die überragende Performance, der allgemeine Support und der Platinum Support deutliche Vorteile, die jeden, der Datenbanken performant, hochverfügbar betreiben will, über Oracle Exadata nachdenken lassen sollten. Lassen Sie sich nicht durch die teils starren vorgegebenen Prozesse durch Oracle (Siteaudit, Auslieferung und Einrichtung) nicht abschrecken. Häufig lassen sich diese durch Beharrlichkeit und Absprechen mit den Oracle Mitarbeitern lösen bzw. anpassen.

Kontaktadresse:
Malthe Griesel
Paragon Data GmbH
Otto-Hahn-Straße 40
D-61381 Friedrichsdorf

Telefon:	+49 (0) 6175-7908330
Fax:	+49 (0) 6175-790861330
E-Mail	M.Griesel@paragon-data.de
Internet:	www.paragon-data.de