Exadata X2: Real-Life-Erfahrungen

Andrzej Rydzanicz, OPITZ CONSULTING GmbH

In diesem Artikel steht der Betrieb der Exadata X2 im Mittelpunkt. Es wird aufgezeigt, dass die Exadata ein äußerst effektives Hilfsmittel zur Erzielung einer höheren Datenbank-Performance ist, das aber trotzdem eine kontinuierliche Überwachung und Wartung erfordert. Anhand von Beispielen kommen darüber hinaus Oracle-Prozeduren zur Sprache, die für den fehlerfreien Betrieb der Maschine notwendig sind.

Bevor man die X-Features der Maschine nutzen kann, muss zunächst der Weg von Oracle bis zum Serverraum des Kunden gebahnt werden. Das hört sich vielleicht banal an – aber die Geschichte kennt bereits Fälle, in denen ganze Projekte an einer zu kleinen Tür auf dem Weg zum Serverraum gescheitert sind. Bevor man also die Exadata bestellt, sollte man die Räumlichkeiten im Rechenzentrum genau unter die Lupe nehmen.

Der Raum muss die erforderliche Größe haben und klimatisiert sein. Die Maschine ist durchaus nicht klein und muss zudem entsprechend den Herstelleranforderungen platziert werden, damit alle Vorgaben für Betrieb und Wartung eingehalten werden (siehe Abbildung 1):

- Tiefe inkl. Türgriff vorn und Türgriff hinten: 1.200 mm
- Abstand von der Rückseite des Racks (Radius der rückwärtigen Tür): 590 mm
- Tiefe ohne Tür: 1.112 mm
- Abstand vom Vorderteil des Racks (Türradius): 638 mm
- Tiefe mit geöffneter Tür: 2.340 mm
- Breite: 600 mm
- Höhe: 1.998 mm

Natürlich kommt die Maschine nicht ohne Verpackung. Was der Kunde letztendlich erhält, ist eine Lieferung per Lkw, die in der Firma durchaus für Verwirrung sorgen kann: Ein Paket, das stolze 1.046,8 kg (Full Rack) wiegt, bekommt man nicht jeden Tag geliefert.

Die Vorbereitung des Serverraums ist das eine, aber das Paket in den Serverraum zu schaffen, das zweite. Jegliche Hindernisse müssen aus dem Weg geräumt werden, damit die Leute, die das rund eine Tonne schwere Paket transportieren, das Gerät auch dorthin stellen können, wo es hingehört. Abmessungen wie Höhe und Breite der Türen (2.184 mm / 1.270 mm), die Fahrstuhltiefe (1.625,6 mm) und die Belastbarkeit des Fahrstuhls (mindestens 1.088 kg) müssen im Vorfeld in Betracht gezogen werden, damit der Weg zum Serverraum nicht zur Qual wird oder sogar Wände eingerissen werden müssen. Neben der Überprüfung potenzieller Hindernisse wie Wände, Türen, Fahrstühle etc. muss der Serverraum selbst entsprechend vorbereitet sein, damit die Maschine korrekt platziert und die Verkabelung an den richtigen Stellen eingesteckt werden kann. Vor allem muss sichergestellt werden, dass die Klimaanlage leistungsfähig genug ist, um das Prachtstück kühlen zu können (siehe Abbildung 2).

Optimal wäre es, die gesamten PDU-Kabel durch Bohrungen im Boden zu führen. Aber keine Sorge: Der Kunde wird bei der Aufstellung vom Hersteller nicht im Stich gelassen. Es gibt diverse Checklisten, die er durchgehen und ausfüllen muss, bevor das große Paket tatsächlich in der Firma eintrifft. Somit ist sichergestellt, dass vor Ort keine unliebsamen Überraschungen auftreten. Es gibt neun unterschiedliche Checklisten, die vor der Bestellung auszufüllen sind:

- Site Readiness (Pass/Conditional Pass/Fail)
- System Components
- Access Route
- DataCenter
- DataCenter Environment
- Facility Power
- Network Configuration and System Software

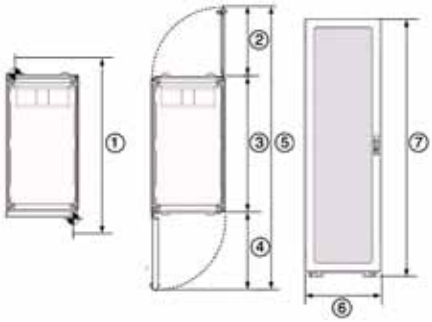

Abbildung 1: Abmessungen des Exadata-Gehäuses

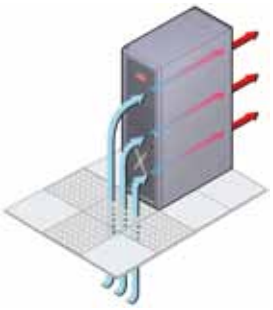

Abbildung 2: Luftbewegungen des Exadata Full Rack im Serverraum

- Logistic
- Safety Items

Erst wenn man alle dort aufgeführten Fragen bearbeitet hat („yes“/„pass“ etc.), kann man sich über die Exadata selbst Gedanken machen. Wer möchte, kann sich die Fragen aus den jeweiligen Listen im „Sun Oracle Database Machine Site Planning Guide“ anschauen – sie sind umfangreich, aber sie gewährleisten einen sicheren Transport und eine korrekte Installation der Maschine am Zielort.

Installation

Nach dem erfolgreichen Transport und Aufbau steht die Exadata nun also endlich im Serverraum und das X-Cabinet

Abbildung 3: Exadata Deployment Assistant

Abbildung 4: Agent Side PSU2

sieht toll aus. Jetzt muss als Erstes ein Oracle-Techniker vorbeikommen und das gute Stück in Betrieb nehmen beziehungsweise installieren. Tatsächlich sind es zwei Techniker, die die Exadata installieren. Einer kontrolliert die Hardware selbst: Er prüft die Netzwerk-Anbindung und aktualisiert die Firmware. Der zweite ist ein Software-Spezialist, der für das Konfigurieren des Real Application Cluster (RAC) verantwortlich ist – er konfiguriert also die SCANs, Listeners, ASM-Diskgroups etc. Die Konfiguration erfolgt mit dem sogenannten „Exadata Deployment Assistant", in dem man die wichtigsten Informationen wie Servernamen, IP-Adressen etc. einträgt (siehe Abbildung 3).

Überwachung und Wartung

Da die Exadata kein Perpetuum Mobile ist, muss sie entsprechend überwacht und gewartet werden, um proaktiv Fehler erkennen zu können. Alle notwendigen Informationen bezüglich der Exadata-Überwachung findet man in My Oracle Support (MOS) unter der Note 1323298.1.

Da die Exadata aus vielen ausgereiften Hardware-Komponenten wie Integrated Lights Out Manager (ILOM), Storage Cells, Infiniband Switches etc. besteht, lohnt es sich aus Erfahrung des Autors nicht, die Maschine durch Open-Source-Mittel wie Nagios zu überwachen, da für die verwendeten Komponenten nur schwer ein Überwachungs-Skript zu finden ist. Natürlich, wenn man Zeit und Geld investieren will, kann man die nötigen Skripte auch selbst schreiben. Es empfiehlt sich eher, Grid Control 11*g* beziehungsweise Cloud Control 12*c* zu nutzen, da für diese Produkte bereits ein Exa-Plug-in existiert, das die wichtigsten Teile der Maschine überwacht beziehungsweise abdeckt.

Bevor man mit dem Plug-in beginnt, müssen zunächst die erforderlichen Grid-Control- oder Cloud-Control-Agenten auf dem jeweiligen Datenbank-Server installiert sein. Oracle verbietet die Installation jeglicher zusätzlicher Software auf den Storage Cells, die Agenten werden also nur auf den Datenbank-Servern installiert und kommunizieren dann mit den Storage Cells via „SSH". Man sollte auch im Hinterkopf behalten, dass der Einsatz des Plug-ins einige zusätzliche Arbeitsschritte auf der Oracle-Management-Server (OMS)- und Agenten-Seite erfordert; OMS- und Agenten-Software sind also zusätzlich zu aktualisieren. Details kann man in der MOS-Note 1323298.1 finden.

Falls jemand zusätzlich alle ORA-Fehler in der Datenbank überwacht und gemeldet bekommen möchte (nicht Exadata-spezifisch, sondern eine Anforderung des Autors), empfiehlt es sich, den 11*g*-Agenten mit PSU2 zu aktualisieren (siehe Abbildung 4). Laut Oracle hat sich das Verhalten ab 11*g* geändert und es werden nur Incidents gemeldet – alles andere wird ignoriert. Im Klartext bedeutet das: Alle, die die „Generic Alert Log"-Metrik im Grid Control vermissen, sind gezwungen, den Agenten PSU2 zu installieren. Danach wird die Metrik wieder im GC 11g sichtbar (Details siehe MOS-Note 8694165.8). Noch eine wichtige Sache: Das Herunterladen des Patch erfordert ein Passwort, das man bei Oracle extra anfordern muss.

Vorort-Service: Exadata-Vorsorgeuntersuchung

Außer einer proaktiven Überwachung via Grid Control 11*g* beziehungsweise Cloud Control 12*c* bietet Oracle einen Vorort-Service zur Funktionsüberprüfung an, um sicherzustellen, dass alle Hardware-Komponenten im laufenden Betrieb korrekt funktionieren. Die Kontrolle mittels Exa-Health-Check beginnt immer am Ende des zweiten Support-Jahres. Der Service wird im Rahmen des Premier Support ausgeführt. Dabei geht es vor allem darum, potenzielle Hardware-Probleme wie defekte Platten (Predictive Failures) oder abgenutzte Storage-Batterien zu erkennen. Der Check prüft ebenfalls diverse Einstellungen auf Datenbank-Ebene (RDBMS, GI, ASM) und schlägt die richtigen Einstellungen vor (Best Practice). Das Bundle „Exadata Database Machine exachk or HealthCheck" kann man aus der MOS-Note 1070954.1 herunterladen. Anschließend entpackt man die *.zip-Datei und führt auf dem Datenbank-Knoten als Oracle User das Kommando „oracle@b-germ-ipxsrv1:~/exacheck[oracle@b-germ-ipxsrv1exacheck]$./exachk" für die jeweilige Datenbank aus. Das Skript fragt nach wichtigen Informationen:

Database Server

Status	Type	Message	Status On	Details
FAIL	ASM Check	ASM processes parameter is not set to recommended value	All ASM Instances	View
FAIL	OS Check	RAID controller battery should be replaced immediately [Database server]	All Database Servers	View
FAIL	OS Check	Database parameter Db_create_online_log_dest_n is not set to recommended value	All Database Servers	View
FAIL	OS Check	Database parameter CLUSTER_INTERCONNECTS is NOT set to the recommended value	All Database Servers	View
FAIL	SQL Parameter Check	Database parameter USE_LARGE_PAGES is NOT set to recommended value	All Instances	View
FAIL	SQL Parameter Check	Database parameter PARALLEL_THREADS_PER_CPU is NOT set to recommended value	All Instances	View
FAIL	OS Check	Database server disk controllers do not use writeback cache	All Database Servers	View
FAIL	OS Check	InfiniBand network error counters are non-zero	All Database Servers	View
FAIL	OS Check	Hardware and firmware profile check not is successful. [Database Server]	All Database Servers	View
FAIL	SQL Check	Some data or temp files are not autoextensible	All Databases	View
FAIL	OS Check	RAID controller battery temperature is not normal [Database Server]	All Database Servers	View
FAIL	ASM Check	ASM Audit file destination file count > 100,000	All ASM Instances	View
FAIL	OS Check	A minimum of two controlfiles are not stored in high redundancy diskgroups	All Database Servers	View
WARNING	OS Check	More than one non-ASM instance discovered	All Database Servers	View
WARNING	SQL Parameter Check	ASM parameter MEMORY_TARGET is NOT set according to recommended value.	All Instances	View
WARNING	SQL Parameter Check	filesystemio_options is not set to recommended value	All Instances	View
INFO	OS Check	ASM griddisk,diskgroup and Failure group mapping not checked.	All Database Servers	View
INFO	SQL Parameter Check	ASM parameter ASM_POWER_LIMIT is set to the default value.	All Instances	View

Abbildung 5: Findings Needing Attention (Database Server) – „exacheck"-Report

Storage Server

Status	Type	Message	Status On	Details
FAIL	Storage Server Check	one or storage server has open critical alerts.	b-germ-ipxstor2	View
FAIL	Storage Server Check	Storage Server alerts are not configured to be sent via email	All Storage Servers	View

Abbildung 6: Findings Needing Attention (Storage Server) – „exacheck"-Report

		Year-end						
		1	2	3	4	5	6	7
Exadata V2	Disk controller battery	No	Yes	No	Yes	No	Yes	No
	Flash ESM	No	No	Yes	No	No	Yes	No
Exadata X2-2, X2-8 and Storage Expansion Rack	Disk controller battery	No	Yes	No	Yes	No	Yes	No
	Flash ESM	No	No	No	Yes	No	No	No

Abbildung 7: Wartung und Austausch von Verschleiß-Komponenten (Consumable Components Maintenance)

	Full System Downtime (batteries replacement only)	Full System Downtime (ESM replacement with or without battery replacement)	Rolling Method
Quarter Rack	1.5 – 2 hours	2 – 2.5 hours	4 hours
Half Rack	2 – 3 hours	2.5 – 4 hours	10 hours
Full Rack	4 – 6 hours	5 – 8 hours	20 hours

Abbildung 8: Proaktive Wartungszeiten (Maintenance Window)

- Name der Datenbank (Welche Datenbank soll kontrolliert werden?)
- Root-Passwort für den Datenbank- und den Storage-Server

In Prinzip ist das alles, was das Skript braucht; alles andere passiert im Hintergrund. Sämtliche relevanten Informationen werden gesammelt und als Endprodukt wird ein HTML-Report generiert, der aus folgenden Kapiteln besteht:

- Findings Needing Attention
- On Database Server
- On Storage Server
- MAA Scorecard
- Findings Passed
- On Database Server
- On Storage Server
- Cluster Wide
- Systemwide firmware and software versions
- Skipped Checks

Der wichtigste Teil ist dabei natürlich „Findings Needing Attention". Hier werden beispielsweise defekte Storage-Batterien angezeigt oder Parameter, die nicht korrekt eingestellt sind. Oracle tauscht Plattencontroller-Batterien (Datenbank-Server und Storage Server) und die sogenannten „Energy Storage Modules" (ESM) in den Flash-Cards natürlich im Rahmen der Garantie aus. Diese Komponenten gehören zu der sogenannten „Consumable-Components-Gruppe", die der Garantie unterliegt. Außerdem wirft ein Oracle-Techniker ein Auge auf das Gehäuse, um defekte Teile zu identifizieren beziehungsweise auszutauschen. Die Abbildungen 5 und 6 zeigen anhand eines Beispiels, wie das Kapitel „Findings Needing Attention" aussehen kann.

Nach Erfahrung des Autors ist es sehr wichtig, die Controller-Batterien im Auge zu behalten, da diese bei Problemen mit der Stromzufuhr die Storage Cell „write cache" schützen. Wenn die Batterie defekt ist, wird die Performance der ganzen Maschine drastisch absinken. Abbildung 7 zeigt, wie oft die Komponenten von Oracle getauscht werden.

Der Austausch der Komponenten kann ohne Downtime erfolgen. Bei einem Plattentausch muss man allerdings im Hinterkopf behalten, dass das ganze System aus dem Backup wiederhergestellt werden muss („Systemwide downtime"), wenn während des Tauschs eine zusätzliche Platte ausfällt („Normal Redundancy"). Bei „High Redundancy" müssen zwei zusätzliche Platten ausfallen. Voraussetzung für den proaktiven Service ist eine Exadata-Software-Version, die 11.2.2.1.1 oder darüber ist. Die V1-Systeme werden darüber nicht abgedeckt (siehe Abbildung 8).

Plattentausch

Wie schon beschrieben, ist die Exadata keine Wundermaschine: Ab und zu fällt eine Platte aus oder eine Batterie muss im Rahmen der Garantie ausgetauscht werden. In den letzten zwei Jahren sind zum Beispiel bei einer Exadata, die die IT-Beratung aus dem Unternehmen des Autors betreut, zwei Platten und alle Controller-Batterien der Storage Cell ausgetauscht worden.

Die Information, dass die Platte defekt ist beziehungsweise bald ausfallen wird („Predicitve Failure"), hat Grid Control 11g geliefert (Exadata Monitoring Plug-in), die Storage-Batterien sind im Rahmen des Exa-Health-Checks ausgetauscht worden. Beide Aktionen erforderten keine Downtime. Sobald das Automatic Storage Management (ASM) feststellt, dass eine Platte ein Problem darstellt, werden die ASM-Disks, die mit der defekten Platte zusammenhängen, automatisch von ASM abgekoppelt und das ASM-Diskgruppen-Rebalancing angestoßen: Die Daten werden von der defekten Platte auf andere ASM-Platten ausgelagert. Das kann eine Weile dauern, daher muss man immer darauf achten, ob das Rebalancing noch im Gange oder schon beendet ist, bevor man mit dem Tausch der Platte beginnt.

Im ersten Schritt öffnet man einen Service Request (SR) bei Oracle, um einen Techniker mit einer Ersatzplatte anzufordern. Natürlich muss das Problem genau beschrieben sein. Im SR selbst wird die Platte bestellt. In der Regel wird das Teil direkt an den Kunden geliefert. Sobald die neue Platte eingetroffen ist, wird automatisch auch der Techniker benachrichtigt, der dann den Kunden anruft und einen Termin für den Einbau vereinbart. Alternativ kann der Techniker selbst die neue Platte direkt mitbringen.

Um den SR korrekt mit den relevanten Informationen zu versorgen, muss zunächst die defekte Platte identifiziert werden. Die erforderlichen Informationen umfassen den Namen der Platte sowie Lun- und Slot-Nummer. Exadata verfügt über ein zusätzliches Storage-Server-Verwaltungs-Tool namens „cellcli". Damit kann man auf Storage-Ebene all diese Informationen auslesen, um anschließend die Service-LED der Platte einzuschalten. Damit weiß der Techniker, welche Platte getauscht werden soll. Der ganze Prozess mit den dazu notwendigen Schritten ist in der MOS-Note 1390836.1 eindeutig erläutert. Der beschriebene Plattentausch-Prozess erfordert keine Downtime und betrifft nur Platten, die für ASM zur Verfügung stehen.

Auch im angesprochenen Fall handelte es sich nur um Platten, die für ASM zur Verfügung standen. Die Prozedur ist komplexer, wenn der Ausfall die Systemplatten betrifft, also Platten, auf denen das Betriebssystem (OS) installiert ist. Aber auch dieser Fall wird in MOS beschrieben.

Batterietausch von Storage Cell und Datenbank-Server

Um die Storage-Controller-Batterie zu tauschen, muss der jeweilige Server heruntergefahren werden. Bei einer der betreuten Exadatas sind beispielsweise alle Batterien getauscht worden, also alle RAID-Batterien für den Datenbank-Knoten und diejenigen für die Storage Cells. Die Entscheidung fiel auf einen Rolling-Austausch, damit keine Downtime entsteht.

Da das Server-Gehäuse abgebaut werden muss, ist auch der jeweilige Knoten herunterzufahren. Bei den Datenbank-Knoten muss man alle Oracle-Dienste, die auf den jeweiligen RAC-Knoten laufen, sauber herunterfahren und dann die Maschine selbst abschalten („#shutdown –h now"). Sobald der Tausch erfolgt ist, kann die Maschine per Knopfdruck wieder eingeschaltet und in Betrieb genommen werden. Danach kann man mit dem nächsten Datenbank-Knoten beginnen. Bei den Storage Cells fällt das ein bisschen komplexer aus. Vor allem muss hier sichergestellt sein, dass die jeweilige Storage Cell ohne Einfluss auf das ASM heruntergefahren werden kann. Alle Schritte sind nachvollziehbar in der MOS-Note 1188080.1 beschrieben. Nachfolgend die wichtigsten Punkte aus der konkret erlebten Praxis.

Da die ASM-Disks nach der Deaktivierung aus ASM entfernt werden, muss sichergestellt sein, dass die Zeit, nach der die Disks abgekoppelt werden, ausreichend lang ist, damit die ASM-Disks nicht vorzeitig deaktiviert werden. Oracle stellt dafür den Parameter „DISK_REPAIR_TIME" zur Verfügung. Damit lässt sich kontrollieren, wann die Disks, die offline sind, aus dem ASM entfernt werden. Standardmäßig ist der Parameter auf 3,6 Stunden eingestellt. Wenn das nicht ausreicht, sollte der Wert auf 8,5 Stunden eingestellt werden. Bevor man mit der Deaktivierung beginnt, ist durch „# cellcli -e list griddisk attributes name,asmmodestatus,asmdeactivationout-com" zu prüfen, ob die Aktion keinen Einfluss auf den Betrieb von ASM beziehungsweise dem ganzen System hat.

Das „asmdeactivationoutcome" sollte für alle Grid-Disks „yes" zurückliefern. Danach deaktiviert man mittels „cellcli -e alter griddisk all inactive" alle Grid-Disks aus der jeweiligen Storage Cell. Anschließend wird die Storage Cell per „#shutdown -h now" heruntergefahren.

Sobald der Techniker alles ausgetauscht und wieder zusammengebaut hat, wird die Storage Cell per Knopfdruck neu gestartet. Am Ende müssen natürlich auch alle Grid-Disks wieder aktiviert werden. Vorher muss allerdings sichergestellt sein, dass die physischen Disks für den Server (Storage Cell) sichtbar sind. Hier sollten sechzehn Geräte (FMODs) und zwölf LSI-Platten angezeigt werden. Die Aktivierung der Grid Disk erfolgt mittels „# cellcli -e alter griddisk all active". Um eine Downtime zu vermeiden, führt man den gesamten Prozess nacheinander für jede Storage Cell durch.

Andrzej Rydzanicz
andrzej.rydzanicz@opitz-consulting.com