Was DBAs über virtualisierte Umgebungen wissen sollten

Siegfried Langer, IBM Deutschland Research & Development GmbH

Server-Virtualisierung ist eine etablierte Methode, um Kosten zu sparen und Ressourcen flexibler zu nutzen. Dies bedingt aber, dass Hardware-Ressourcen gemeinsam genutzt werden und einzelne Server sich in diese „Gemeinschaft von Servern" einfügen müssen.

Aus der Server-Virtualisierung ergeben sich neue Herausforderungen an das Kapazitäts- und Performance-Management. Das Tuning einzelner Anwendungen kann Auswirkungen auf andere Anwendungen in der virtuellen Umgebung haben und ein Systemadministrator muss sicherstellen, dass eine faire Zuteilung von Ressourcen erfolgt, gleichzeitig aber auch Service Level Agreements (SLA) eingehalten werden können. Dieser gesteigerten Komplexität stehen allerdings auch Vorteile gegenüber: Neben Kosteneinsparungen durch effizientere Nutzung der Infrastruktur wird der individuelle Wartungsaufwand für einzelne, jetzt virtuelle Server geringer und das Betriebsmanagement – wie Backup und Vorsorge für den Katastrophenfall – lässt sich einheitlicher gestalten.

Für den DBA ist es wichtig zu verstehen, dass Performance-Probleme aufgelöst werden sollten, anstatt sie mit mehr Ressourcen (Cores, Speicher) herunterzuspielen, was zu erheblichen Mehrkosten führen kann (etwa durch Software-Lizenzkosten). In einer virtuellen Umgebung kann „mehr" (Aufwand) manchmal sogar „weniger" (Performance) sein.

Einzelne physische Server müssen für Lastspitzen ausgelegt sein. Die erforderliche Leistung muss die erwartete Spitzenauslastung sowie zukünftige Wachstumsreserven berücksichtigen, auch wenn diese nur kurzzeitig auftreten. Aufgrund der relativ geringen Hardwarekosten und der Verfügbarkeit von Multi-Core-Servern stellt dies meist kein unmittelbares Problem dar.

Wenn man allerdings die Software-Lizenzkosten einbezieht, die sich typischerweise an der Anzahl der Prozessorkerne (Cores) bemessen, ergibt sich ein anderes Bild. Die Konsolidierung vieler, teils nur wenig ausgelasteter Prozessoren auf einen hoch virtualisierten Server führt zu einer wesentlich besseren Nutzung der Ressourcen und ermöglicht erhebliche Einsparungen bei den Software-Lizenzkosten. Darüber hinaus ergeben sich teils erhebliche Einsparungspotenziale bei den operativen Kosten (Strom, Kühlung, Stellfläche, Netzwerk, Servicepersonal). Ein zentralisiertes Management reduziert den Verwaltungsaufwand und erlaubt zentralisierte Datensicherung, bessere Vorsorge für den Katastrophenfall, Hochverfügbarkeit sowie die Nutzung von Cloud-Konzepten mit hoher Flexibilität und schneller Aktivierung neuer Server. Virtualisierung ist ein wich-

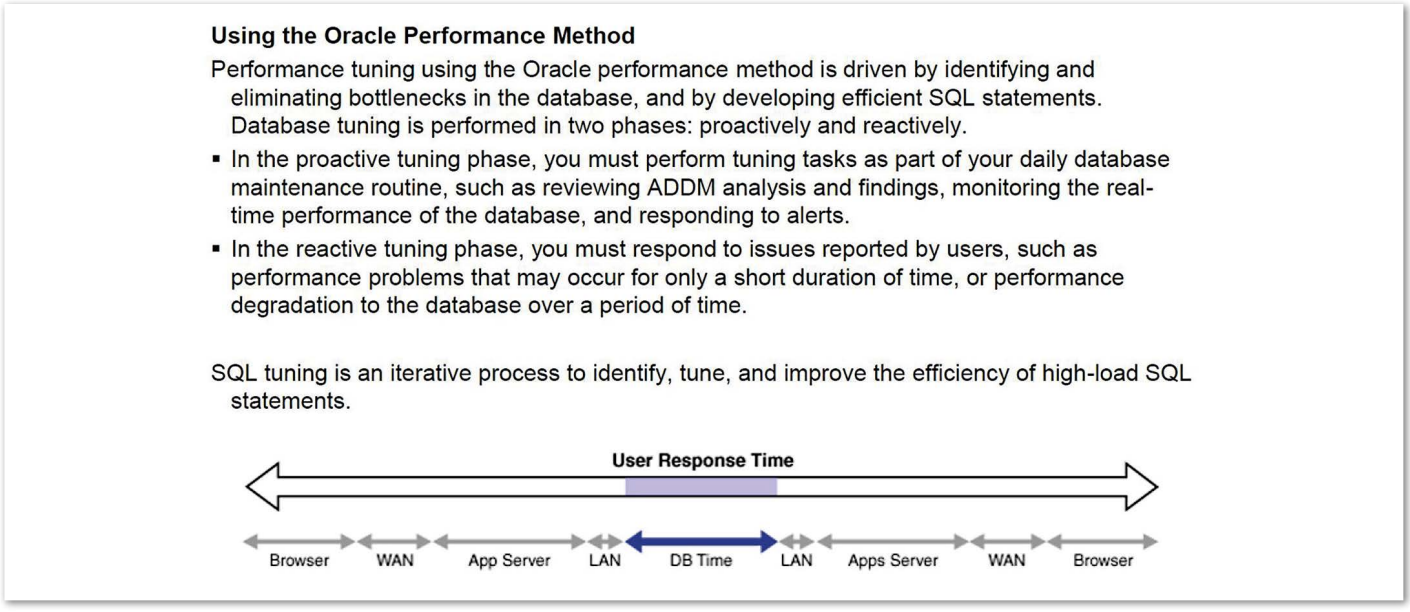


Abbildung 1: Auszug aus dem Oracle Tuning Guide

tiger erster Schritt, um IT-Ressourcen für Cloud-Services nutzen zu können. Nur so kann ein hochflexibler und kosteneffektiver Betrieb gewährleistet werden.

Performance-Management

Das Tuning eines Systems sollte von gemessenen Daten (Baseline) ausgehen und als sich ständig wiederholender Kreislauf von „Messen" – „Evaluieren" – „Verbessern (Verändern)" – „Messen" betrachtet werden. Veränderungen sollten sich auf einen oder wenige Parameter beziehen, da verschiedene Tuningmaßnahmen sich gegenseitig beeinflussen und sogar kompensieren können *(siehe Abbildung 1, Quelle „http://docs.oracle.com/cd/E24628_01/server.121/e17635/tdppt_method.htm#TDPPT006")*.

Endbenutzer beurteilen den Durchsatz typischerweise über die beobachtete Antwortzeit. Hier ist zu beachten, dass die Datenbank-Zeit nur einen Teil der Antwortzeit ausmacht. Die Ursache für solche Probleme kann auch in der Anwendung oder im Netzwerk liegen.

Häufig ist zu beobachten, dass Performance-Probleme durch mehr Ressourcen adressiert werden. Natürlich lässt sich eine ineffektive Anwendung häufig durch mehr Prozessoren und/oder Speicher „erschlagen", aber es werden in diesem Fall nur die Symptome beseitigt, die Ursache bleibt bestehen. Eine solche Vorgehensweise kann schnell zu einer Kostenexplosion und möglicherweise zu komplexen Folgefehlern führen.

Virtualisierte Umgebungen

Virtualisierte Server teilen sich physische Ressourcen. Dies hilft bei Lastspitzen, es ist jedoch kein Rezept, um Performance-Engpässe zu beseitigen. Da nun mehrere Server um die physischen Ressourcen konkurrieren, kann die Suche nach solchen Flaschenhälsen wesentlich komplizierter werden, insbesondere wenn einzelne Server nicht optimal konfiguriert sind. Es gelten folgende Grundsätze:

Definiere nicht mehr virtuelle CPUs für einen Linux-Gast als nötig:

- Die Nutzung mehrerer Prozessoren benötigt Software-Locks, sodass Daten oder Kontrollblöcke nur von einem Prozessor zu einer Zeit geändert werden können.
- Linux nutzt ein globales Lock. Wenn das Lock gehalten wird und ein anderer Prozessor es benötigt, muss er warten.
- Die Zahl der virtuellen Prozessoren sollte nach dem Bedarf gesetzt werden und nicht einfach der Anzahl der realen Prozessoren entsprechen.
- Vorsicht beim Klonen: Einige Linux-Gäste brauchen mehr virtuelle CPUs als andere, etwa Oracle-Datenbankserver.

Definiere den (virtuellen) Speicher für Linux nicht größer als nötig:

- Exzessive virtuelle Speichergrößen haben eine negative Auswirkung auf die Performance.
- Linux nutzt freien Speicher für das Caching von Daten. Für gemeinsam genutzte (shared) Ressourcen hat dies negative Auswirkungen.

- Reduziere die Größe des Linux-Gastes, bis er beginnt, Speicher auszulagern (Swap).
- Benutze virtuelle Plattenspeicher (VDISK) für Swap, wenn genügend realer Speicher verfügbar ist.
- Vergleiche die Linux-Speichernutzung mit den im Hypervisor definierten Größen des Gastes.

Oracle-Datenbanken auf VMware

VMware gibt Empfehlungen im „Oracle Databases on VMware Best Practices Guide" *(siehe „http://www.vmware.com/files/pdf/solutions/oracle/Oracle_Databases_VMware_Best_Practices_Guide.pdf")*. Diese haben durchaus allgemeingültigen Charakter:

- *Definiere so wenig virtuelle Prozessoren (vCPUs) wie möglich*
  Sofern das Monitoring der aktuellen Arbeitslast keine Verbesserung des Durchsatzes der Oracle-Datenbank durch mehr virtuelle Prozessoren zeigt, führen die zusätzlichen vCPUs zu Engpässen im Scheduler und können die Gesamt-Performance des virtualisierten Servers beeinträchtigen.
- *Speicher-Reservierungen sollten gleich der Oracle-SGA-Größe gesetzt sein*
  Der reservierte Speicher sollte groß genug sein, um Speicherauslagerungen (kernel swapping) zwischen ESX und den Gastbetriebssystemen zu vermeiden.
- *Nutze Oracle Automatic Storage Management (ASM)*
  ASM bietet integriertes Cluster File System und Platten (Volume) Management für Oracle-Datensätze. Es vereinfacht das Anlegen von Datensätzen und ermöglicht einen Daten-Durchsatz, der nahe an die Roh-Datenrate der Plateneinheiten herankommt.
- *Folge den „Best Practices"-Empfehlungen der Plattenhersteller beim Anlegen von Oracle-Datenbanken*
  ASM ist nicht in der Lage, die optimale Platzierung der Daten oder die LUN-Auswahl für das verwendete Plattenspeichersystem zu bestimmen. Daher ist es kein Ersatz für die enge Abstimmung zwischen Plattenspeicher- und Datenbank-Administratoren *(siehe Abbildung 2)*.
- *ASM-Plattenspeicher-Gruppen (disk groups) sollten gleiche Plattentypen mit gleicher Geometrie beinhalten*
  Es sollten mehrere ASM-Disk-Gruppen basierend auf den I/O-Charakteristika angelegt werden. Minimum sind zwei ASM-Disk-Gruppen, eine für Log-Files, die sequenzieller Natur sind, und eine andere für Daten, die von Natur aus wahllos (random) sind. Für hohe Durchsatzraten wird empfohlen, mehrere parallele Datenpfade zu den Datenplatten zu definieren beziehungsweise die Daten auf mehrere ASM-Gruppen zu verteilen.

Oracle-Datenbank-Konsolidierung auf dem IBM System z

„IBM System z" heißt die Produktfamilie für den IBM-Mainframe. Neben den traditionellen Mainframe-Betriebssystemen wie z/OS oder z/VSE ist auch Linux unterstützt und nutzt spezielle System-z-Prozessoren, die Integrated Facility for Linux (IFL). Die System-z-Architektur unterstützt zwei Virtualisierungsebenen: Das physische System kann in bis zu sechzig logische Partitionen (LPARs) aufgeteilt sein, wobei Prozessoren mehrfach genutzt werden können (shared). Diese erste Ebene bietet eine sehr hohe Isolation zwischen den Partitionen, die nach Common Criteria „EAL5" zertifiziert ist, was als äquivalent zu physisch getrennten Systemen gilt. Dadurch ist es möglich, Produktions-, Test- oder Entwicklungssysteme gleichzeitig auf dem gleichen Rechner zu betreiben ohne das Risiko, dass Abstürze im Testbetrieb zu Ausfällen des Produktionssystems führen. Eine weitere, wesentlich granularere und flexiblere Virtualisierungsebene bietet der Hypervisor z/VM.

Der IBM System z Hypervisor z/VM

„Linux on System z" kann sowohl „bare metal" im LPAR laufen als auch unter z/VM. Als Besonderheit kann z/VM auch die traditionellen Betriebssysteme virtualisieren. Es ist sogar möglich, den z/VM Hypervisor unter z/VM zu betreiben, was beispielsweise für Schulungszwecke gerne genutzt wird. z/VM bietet höchste Skalierbarkeit für eine virtuelle Serverumgebung durch die Kombination von virtuellen und realen Ressourcen für jede virtuelle Maschine.

Der Hypervisor z/VM erlaubt es, den einzelnen Gastsystemen mehr Speicher zuzuweisen als physisch vorhanden (memory overcommitment). Dies ist eine sehr praktische Funktion bei vielen Gästen, die nur geringe oder seltene Anforderungen stellen (etwa Test- oder Entwicklungssysteme), da dieser Speicher den aktiven Prozessen zur Verfügung gestellt werden kann.

Überdimensionierte Gäste binden wertvolle Ressourcen, die der Hypervisor im Gesamtsystem mühsam suchen müsste. Außerdem ist die Zeit, die für das Verschie-

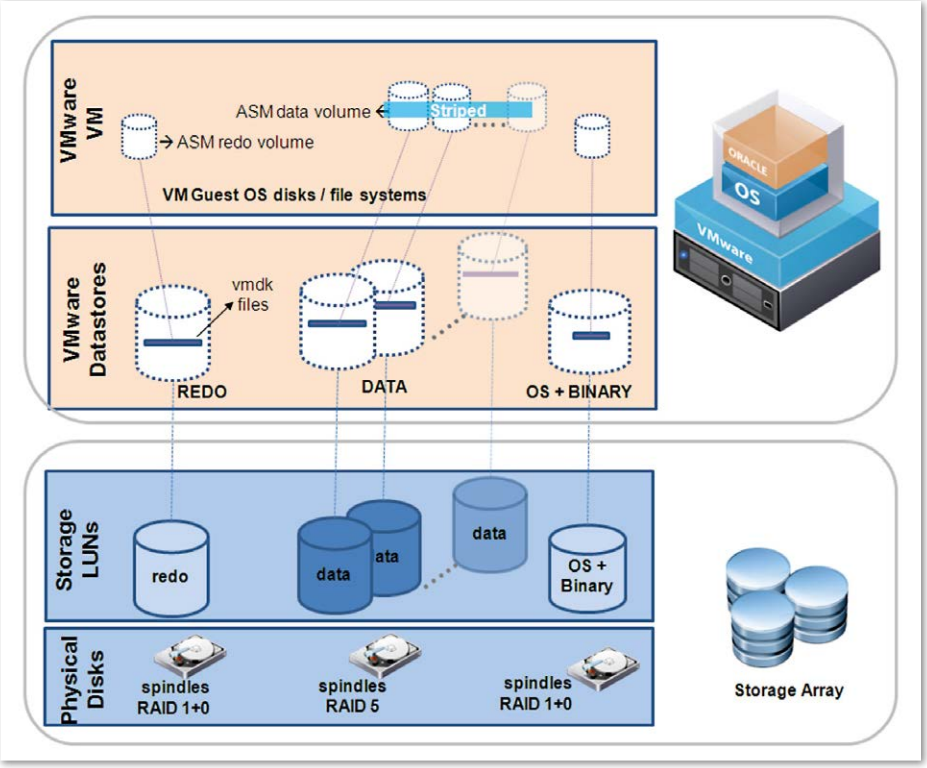


Abbildung 2: Speicher-Layout-Beispiel für OLTP-Datenbank mit VMware

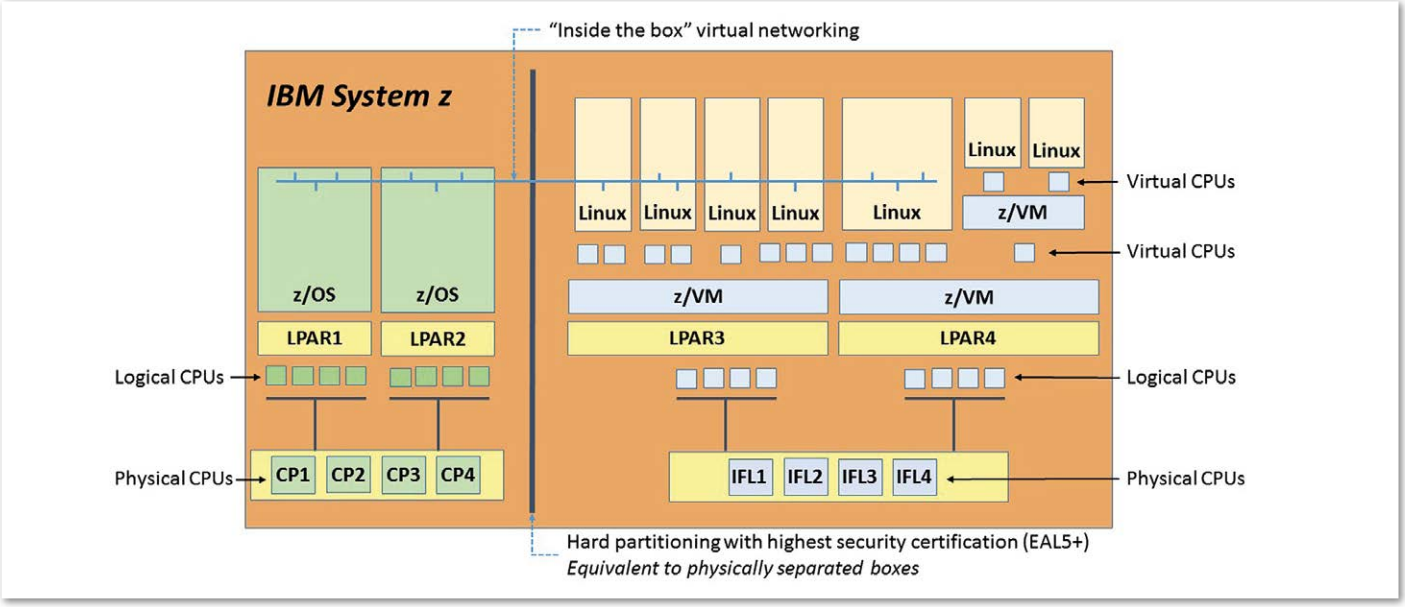


Abbildung 3: Extreme Virtualisierung mit z/VM

ben von Gigabyte-großen Speicherinhalten notwendig ist, unter Durchsatzgesichtspunkten nicht zu vernachlässigen. Für Linux-Gäste gilt daher, dass diese nur so groß dimensioniert werden sollten, wie für eine gute Funktionalität notwendig ist.

Zusätzliches Caching im Linux-Gast, insbesondere I/O-Caching mit laufenden Updates, bindet wertvollen Speicher, der anderen Prozessen nicht zur Verfügung steht. Da der Hypervisor in erster Linie nach dem „least recently used"-Algorithmus vorgeht, wird dieser Cache-Speicher nicht angefasst. Die empfohlene Vorgehensweise ist hier, Direct-I/O zu verwenden und die Linux-Gäste speicherseitig so zu dimensionieren, dass das individuelle Linux im Normalbetrieb gerade noch nicht auf den externen Speicher auslagert (swapped). Dies stellt sicher, dass z/VM das Gesamtsystem optimal mit Ressourcen versorgen kann. Für die Speicher-Dimensionierung hat sich die Faustformel „Startgröße = SGA + PGA + Linux + ASM" für die virtuelle Linux-Server-Speichergröße bewährt:

- Speicher für Oracle SGA und PGA: laut DBA-Abschätzung
- Speicher für Linux Kernel: 512 MB
- Speicher für Oracle ASM: 256 MB bis 512 MB (falls ASM verwendet wird)

Das Verhältnis zwischen virtueller Speichergröße (Summe der definierten Speichergrößen aller virtuellen Server des Hypervisor) und realem Speicher, also „Virtual:Real", sollte kleiner als „3:1" sein. Dieser Wert ist durchaus praxisrelevant für Test- und Entwicklungsumgebungen. Für Oracle-Datenbanken im Produktionsbetrieb hat sich ein Startpunkt von „1,5:1" als brauchbarer Kompromiss erwiesen. Für besonders performancekritische Produktionsanwendungen kann es sinnvoll sein, das Verhältnis bis auf „1:1" zu ändern.

Ähnliche Überlegungen gelten auch für die Zuordnung physischer und virtueller Prozessoren (CPU). Auch hier führt eine Überdimensionierung zu mehr Verwaltungsaufwand und damit Overhead im Hypervisor. Zusätzlich besteht das Risiko, dass sehr CPU-aktive Prozesse das Gesamtsystem dominieren und andere Gäste nur noch unzureichend Service geben können, da sie nicht mehr die benötigten Prozessor-Zeitscheiben zugewiesen bekommen. Die vorab zitierten Empfehlungen, bezogen auf virtuelle CPU, Plattenspeicher und ASM, die für Oracle-Datenbanken auf VMware gegeben wurden, gelten prinzipiell auch für z/VM.

Automatisierung des Resource Managements

Die Größenbestimmung von Linux-Gästen kann ein komplexes Unterfangen sein, insbesondere in einem dynamischen Umfeld mit wechselnden Anforderungen und schnell wachsenden Anwendungen. Zu groß dimensionierte Linux-Gäste kosten zusätzlichen Management-Aufwand im Hypervisor, zu klein dimensionierte Gäste führen zu Performance-Problemen, insbesondere in Auslastungsspitzen.

Es besteht die Möglichkeit, das Management der Ressourcen zu automatisieren. Basierend auf den Anforderungen des Gastes kann das System CPUs und Speicher nach vordefinierten Regeln hinzufügen oder entfernen. Diese Funktion wird durch den Linux „cpuplugd daemon" (auch „hotplug daemon" genannt) zur Verfügung gestellt und ist für Linux on System z ab SLES 11 SP2 oder RHEL 6.2 verfügbar *(siehe „http://www-01.ibm.com/support/knowledgecenter/linuxonibm/liaag/l0cpup00_2012.htm?cp=linuxonibm%2F0-4-3-1-2")*.

Fazit

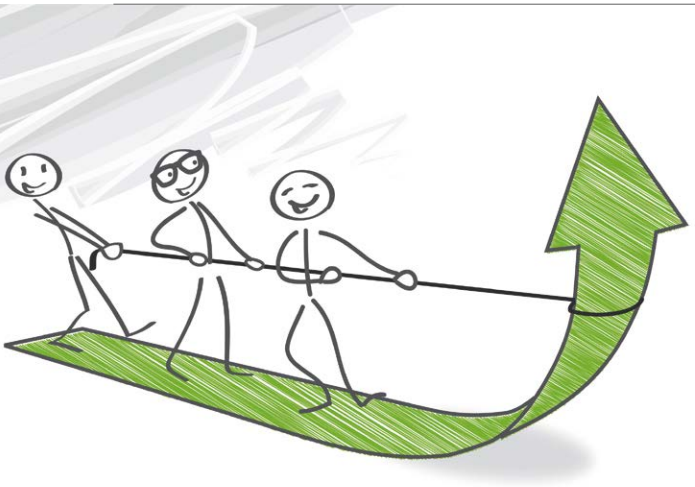
Die Befolgung der genannten Hinweise und Empfehlungen allein garantiert noch nicht, dass eine virtualisierte Umgebung alle Anforderungen der Benutzer erfüllt, sie macht es jedoch einfacher, Ursachen zu ergründen und Abhilfe zu schaffen. Zusammenfassend soll an die Basisregeln des Performance-Managements erinnert werden:

- Etabliere ein permanentes Monitoring

- Sammle Systemdaten als Basisbewertung für gute Performance
- Implementiere einen Change-Management-Prozess
- Führe zu einer Zeit so wenige Änderungen wie möglich durch
- Performance ist oft nur so gut wie das schwächste Glied
- Die Beseitigung eines Flaschenhalses führt zu weiteren neuen Flaschenhälsen
- Erwarte Veränderungen an anderer Stelle, wenn eine Ressource verändert wird

Grundsätzlich gilt, dass die Ursachenforschung bei Durchsatzproblemen am Anfang stehen sollte. Das beste Tuning von System-Ressourcen kann Probleme der Anwendung oder der Netzwerk-Anbindung nicht beseitigen.

Siegfried Langer
siegfried.langer@de.ibm.com



Aus der Exadata-Konsolidierung wird eine Oracle Engineered Architecture

Christian Trieb, Paragon Data GmbH

Nach einer mehr oder weniger längeren Zeitspanne sind viele Datenbanken auf einer Exadata-Datenbankmaschine konsolidiert, das System ist ausgelastet und es gibt keine Datenbanken mehr, die auf die Exadata konsolidiert werden sollen. Letzteres ist sicher selten, kann aber (theoretisch) vorkommen. Dieser Artikel betrachtet den Betrieb einer Exadata-Maschine und die aus den anstehenden Anforderungen resultierenden Weiterentwicklungen.

Betrieb

Im Rechenzentrumsbetrieb stellt sich eine Exadata auf den ersten Blick wie ein normales Server-Rack dar. Allerdings merkt man schon in der Planung eines Exadata-Betriebs schnell, dass doch einiges zu beachten ist. So muss bei der Integration und dem Betrieb einer Exadata in einem Rechenzentrum die Vorgehensweise gut geplant sein. Mehrere Abteilungen (Betriebssystem-, Storage-, Netzwerk- und Datenbank-Administration) müssen diese gemeinsam sicherstellen. Das klassische Denken in Silos muss also aufgebrochen werden. Nur so lassen sich Synergie-Effekte nutzen. Im Unternehmen des Autors wurde dies unter der Leitung der Datenbankadministratoren gut umgesetzt. Die Verantwortung für die komplette Maschine liegt daher bei den DBAs. Sie wurden weitergebildet und können selbstständig die meisten Aufgaben lösen. Bei komplexeren Fragestellungen und Detailproblemen werden die Kollegen der anderen Abteilungen hinzugezogen.

Für den Betrieb einer Exadata-Maschine kommen die gleichen Methoden und Werkzeuge wie für den Betrieb von anderen Servern und Datenbanken zum Einsatz. Die Maschine wird regelmäßig gesichert und mit dem Tool „Nagios“ überwacht. Datenbanken werden mit RMAN gesichert, auch mit „Nagios“ überwacht und mit Oracle Enterprise Manager Cloud Control administriert. Das Patchen einer Exadata-Maschine funktioniert mithilfe des Oracle Platinum Supports und einer sehr langfristigen detaillierten Planung relativ gut.

Anforderungen

Im Laufe des Betriebs der Exadata-Maschine (X3-2 Quarter Rack) stellte sich nach Integration von immer mehr Datenbanken (mehr