Was DBAs über virtualisierte Umgebungen wissen sollten

Siegfried Langer
IBM Deutschland Research & Development GmbH
Böblingen

Schlüsselworte

Oracle DB, Konsolidierung, Virtualisierung, Optimierung, Performance Tuning, Linux, IBM Linux on System z, Mainframe, z/VM Hypervisor, Best Practices, VMware.

Einleitung

Server-Virtualisierung ist eine etablierte Methode, um Kosten zu sparen und Ressourcen flexibler nutzen zu können. Dies bedingt aber, dass Hardwareressourcen nun gemeinsam genutzt werden und einzelne Server sich in diese "Gemeinschaft von Servern" einfügen müssen. Daraus ergeben sich neue Herausforderungen an das Kapazitäts- und Performancemanagement. 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 Servicelevel-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, das Betriebsmanagement, wie beispielsweise Backup und Vorsorge für den Katastrophenfall, kann einheitlicher gestaltet werden.

Für den DBA ist es wichtig zu verstehen, dass Performanceprobleme gelöst werden sollten, anstatt sie mit mehr Ressourcen (mehr Cores, mehr Speicher) herunterzuspielen (was zu erheblichen Mehrkosten führen kann - z.B. Softwarelizenzkosten). In einer virtuellen Umgebung kann "mehr" manchmal sogar "weniger" [Performance] sein.

Es werden die grundsätzlichen Unterschiede zwischen physischen Einzelsystemen und virtualisierten Umgebungen betrachtet, die allgemeingültig sind und für alle virtualisierten Serversysteme, wie VMware, XEN, KVM, HyperV, Oracle VM Server, z/VM, gelten. Praktische Beispiele werden anhand von Oracle DB auf einer hoch-virtualisierten Linux on System z Umgebung mit z/VM erläutert.

Virtualisierung – die Voraussetzung für eine optimierte und konsolidierte Systemumgebung

Einzelne physische Server müssen für Lastspitzen ausgelegt werden. Die erforderliche Leistung muss die erwartete Spitzenauslastung und 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 aber die Software-Lizenzkosten, die sich typischerweise an der Anzahl der Prozessorkerne (Cores) bemessen, einbezieht, dann 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 Einsparungspotentiale 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 und die Nutzung von Cloudkonzepten mit hoher Flexibilität und schneller Aktivierung neuer Server.

Virtualisierung ist ein wichtiger erster Schritt, um IT Ressourcen für Cloud-Services nutzen zu können. Nur so kann ein hoch-flexibler und kosteneffektiver Betrieb gewährleistet werden.

Grundsätzliches zum Performance-Management

Wikipedia definiert Leistungsmanagement oder Performance-Management (unter anderem) mit folgenden Worten: „Zielsetzung der Ansätze des Performance-Managements ist eine systematische, mehrdimensionale Leistungsmessung, -steuerung und -kontrolle … mit dem Ziel der kontinuierlichen Verbesserung von individueller und Unternehmensleistung."

Das Tuning eines Systems sollte von gemessenen Daten (Baseline) ausgehen und als sich ständig wiederholender Prozess 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.

Endbenutzer beurteilen den Durchsatz typischerweise über die beobachtete Antwortzeit. Hierbei 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.

Using the Oracle Performance Method

Performance tuning using the Oracle performance method is driven by identifying and eliminating bottlenecks in the database, and by developing efficient SQL statements. Database tuning is performed in two phases: proactively and reactively.

- In the proactive tuning phase, you must perform tuning tasks as part of your daily database maintenance routine, such as reviewing ADDM analysis and findings, monitoring the real-time performance of the database, and responding to alerts.
- In the reactive tuning phase, you must respond to issues reported by users, such as performance problems that may occur for only a short duration of time, or performance degradation to the database over a period of time.

SQL tuning is an iterative process to identify, tune, and improve the efficiency of high-load SQL statements.

Abbildung 1 Auszug aus dem Oracle Tuning Guide

Häufig ist zu beobachten, dass Performanceprobleme durch mehr Ressourcen adressiert werden. Natürlich kann eine ineffektive Anwendung oder suboptimale Datenstruktur häufig durch mehr Prozessoren und/oder Speicher „erschlagen" werden, aber es werden in diesem Fall nur die Symptome beseitigt, die Ursache bleibt bestehen und eine solche Lösungsweise führt schnell zur Kostenexplosion.

Was ist anders in virtualisierten Umgebungen

Virtualisierte Server teilen sich physische Ressourcen. Dies hilft bei Lastspitzen, aber es ist kein Rezept, um Performance-Flaschenhälse 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. Insbesondere in virtuellen Umgebungen kann „mehr" oft „weniger" bedeuten.

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 Clonen: einige Linux-Gäste brauchen mehr virtuelle CPUs als andere, z.B. Oracle Datenbankserver.

Definiere den (virtuellen) Speicher des 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).
- 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“. Diese Empfehlungen 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 Gesamtperformance des virtualisierten Servers beeinträchtigen.

Speicherreservierungen sollten gleich der Oracle SGA Größe gesetzt werden!

Der reservierte Speicher sollte groß genug sein, um Speicherauslagerungen (kernel swapping) zwischen ESX und den Gastbetriebssystemen zu vermeiden.

Nutze Oracle Automatic Storage Management (ASM)!

Oracle ASM bietet integriertes Cluster File System und Platten (Volume) Management für Oracle Datensätze. ASM vereinfacht das Anlegen von Datensätzen und ermöglicht einen Daten-Durchsatz, der nahe an die Roh-Datenrate der Platteneinheiten herankommt.

Folge den „Best Practices“ Empfehlungen der Plattenhersteller beim Anlegen von Oracle Datenbanken!

Oracle ASM ist nicht in der Lage, die optimale Platzierung der Daten oder LUN Auswahl für das verwendete Plattenspeichersystem zu bestimmen. Daher ist Oracle ASM kein Ersatz für die enge Abstimmung zwischen Plattenspeicher- und Datenbank-Administratoren.

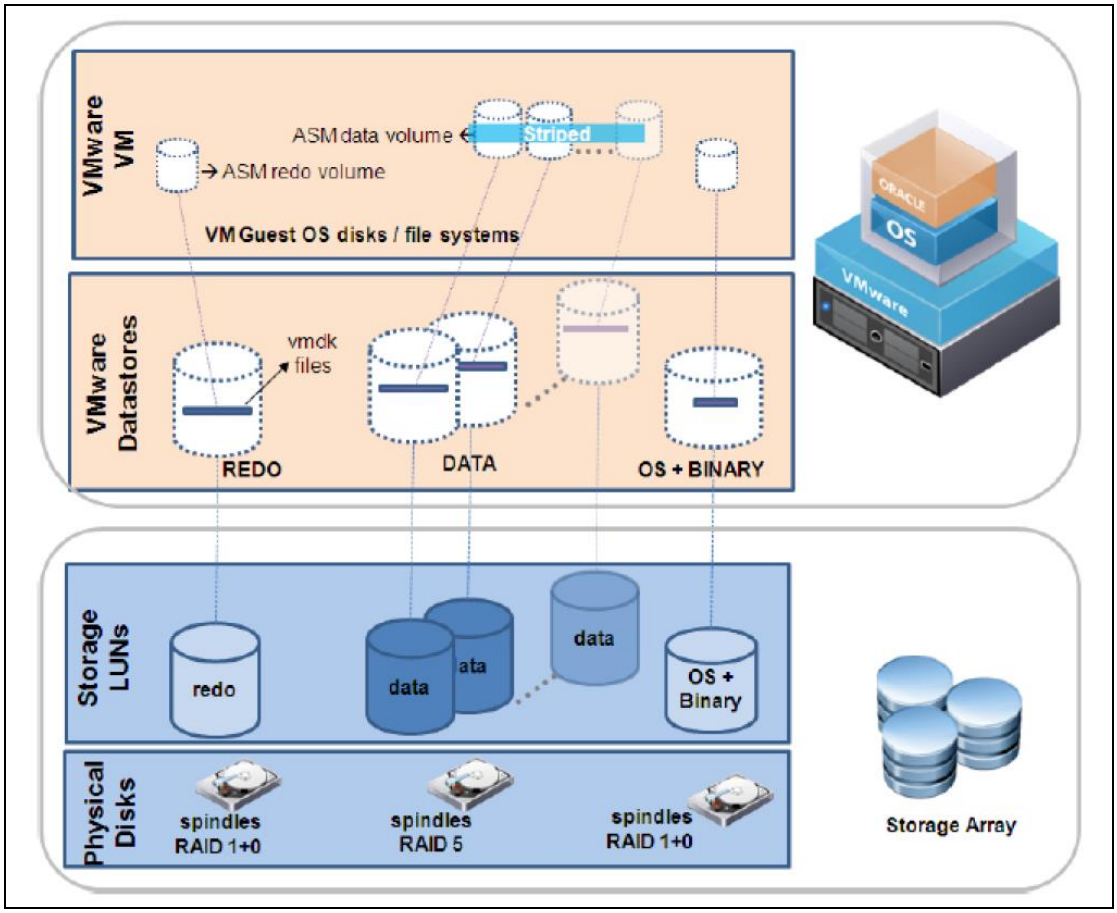

Abbildung 2 Speicher-Layout - Beispiel für OLTP Datenbank mit VMware (Quelle: Oracle Databases on VMware Best Practices Guide)

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. Als Minimum sollten zwei ASM Disk Gruppen angelegt werden: eine für Log-Files, welche sequentieller 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, bzw. die Daten auf mehrere ASM Gruppen zu verteilen.

Der IBM System z Hypervisor z/VM

„IBM System z“ ist der Produktfamilienname für den IBM Mainframe. Neben den traditionellen Mainframe-Betriebssystemen, wir 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 60 logische Partitionen (LPARs) aufgeteilt werden, 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- und 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 granulare und flexiblere Virtualisierungsebene bietet der 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.

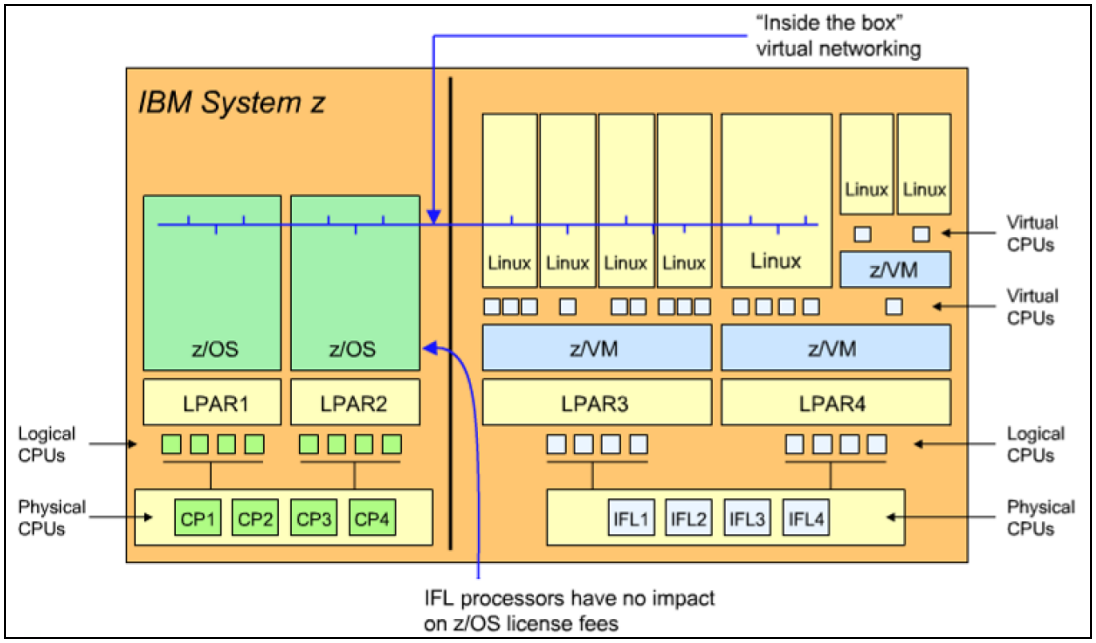

Abbildung 3 Extreme Virtualisierung mit z/VM

Der Hypervisor z/VM erlaubt es, den einzelnen Gastsystemen mehr Speicher, als physisch vorhanden, zuzuweisen (memory overcommitment). Dies ist eine sehr praktische Funktion bei vielen Gästen, die nur geringe oder seltene Anforderungen stellen (z.B. 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 Verschieben von gigabytegroßen Speicherinhalten notwendig ist, unter Durchsatzgesichtspunkten nicht zu vernachlässigen. Für Linux on System z Gäste gilt daher, dass diese Linux-Gäste nur so groß dimensioniert werden sollten, wie für eine gute Funktionalität notwendig.

Zusätzliches Caching im Linuxgast, 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 Cachespeicher nicht angefasst. Die empfohlene Vorgehensweise ist hier, Direct I/O zu verwenden und die Linuxgäste speicherseitig so zu dimensionieren, dass das individuelle Linux im Normalbetrieb gerade noch nicht auf den externen Speicher auslagert (swapped). Dadurch wird sichergestellt, dass z/VM das Gesamtsystem optimal mit Ressourcen versorgen kann.

Folgende Speicherdimensionierung hat sich als Faustformel für die virtuelle Linux-Server Speichergröße bewährt:

```
Startgröße = SGA + PGA + Linux + ASM
```

- 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 Hypervisors) und realem Speicher „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 bewiesen. Für besonders performancekritische Produktionsanwendungen kann es sinnvoll sein, das Verhältnis auf 1:1 zu bringen.

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 Gästen 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 Ressource Managements

Die Größenbestimmung von Linux Gästen unter z/VM kann ein komplexes Unterfangen sein, insbesondere in einem dynamischen Umfeld mit wechselnden Anforderungen und schnell wachsenden Anwendungen. Zu groß dimensionierte Linux-Gäste kosten zusätzlichen Managementaufwand im Hypervisor, zu klein dimensionierte Gäste führen zu Performanceproblemen, 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. Ein Papier mit weitere Informationen und Testergebnissen ist im Anhang referiert.

Eine weitere Möglichkeit die Ressourcen den dynamischen Unternehmensanforderungen anzupassen, wird durch die System z „On Demand Capacity" Option geboten. On/Off Capacity on Demand (CoD) erlaubt es, temporär weitere Prozessoren (IFLs) zu aktivieren, um Belastungsspitzen abzufangen. Dies kann dynamisch während des Betriebs erfolgen, setzt allerdings nicht genutzte Prozessoren in der Hardwareausstattung voraus.

Eine weitere Option, um die Kapazität temporär zu erhöhen, ist Capacity Backup Upgrade (CBU). CBU erlaubt es, nicht aktivierte Prozessoren für einen begrenzten Zeitraum zu aktivieren, um Kapazität von einem Betriebsrechner auf einen anderen Rechner im Unternehmen zu verlagern. Typischerweise wird CBU eingesetzt, wenn ein Rechner ausfällt oder aufgrund eines Katastrophenfalls nicht genutzt werden kann. Der Reserverechner kann im Normalbetrieb mit geringerer Kapazität laufen und im Katastrophenfall kurzfristig in seiner Kapazität vergrößert werden, was zu erheblichen Kosteneinsparungen führen kann.

Zusammenfassung

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

- Etablieren Sie ein permanentes Monitoring.
- Sammeln Sie Systemdaten als Basisbewertung für gute Performance.
- Implementieren Sie einen Change-Management-Prozess.
- Führen Sie so wenige Änderungen wie möglich zu einer Zeit durch.
- Performance ist oft nur so gut, wie das schwächste Glied.
- Die Beseitigung eines Flaschenhalses führt zu weiteren, neuen Flaschenhälsen.
- Erwarten Sie Veränderungen an anderer Stelle, wenn eine Ressource verändert wird.

Kontaktadresse:

Siegfried Langer

IBM Deutschland Research & Development GmbH

Schönaicher Straße 220

D-71032 Böblingen

Telefon: +49 (0) 7031-16 4228

Fax: +49 (0) 7031-16 3456

E-Mail: Siegfried.Langer@de.ibm.com

Internet: www.ibm.com

Weitere Informationen:

Oracle:

Oracle® Database 2 Day + Performance Tuning Guide, 12c Release 1 (12.1), E17635-10

http://docs.oracle.com/cd/E24628_01/server.121/e17635/tdppt_method.htm#TDPPT006

VMware:

Oracle Databases on VMware Best Practices Guide

http://www.vmware.com/files/pdf/solutions/oracle/Oracle_Databases_VMware_Best_Practices_Guide.pdf

IBM:

Using the Linux cpuplugd Daemon to manage CPU and memory resources from z/VM Linux guests

http://www-01.ibm.com/support/knowledgecenter/linuxonibm/liaag/l0cpup00_2012.htm?cp=linuxonibm%2F0-4-3-1-2

IBM Redbooks:

Experiences with Oracle Database 12c Release 1 on Linux on System z

http://publib-b.boulder.ibm.com/abstracts/sg248159.html?Open

Experiences with Oracle 11gR2 on Linux on System z

http://publib-b.boulder.ibm.com/abstracts/sg248104.html?Open

Installing Oracle 11gR2 RAC on Linux on System z

http://www.redbooks.ibm.com/abstracts/redp4788.html?Open