



Oracle Datenbank-Serverkonsolidierung mit Linux on System z

Ein Erfahrungsbericht

Siegfried Langer

IBM Deutschland Research & Development



I'm
Speaker

20. - 22.11.2012
Nürnberg

2012
DOAG
Konferenz + Ausstellung

DOAG Vortrag Nr. 95



© 2012 IBM Corporation

Diskussionspunkte

Konsolidierung von Oracle Datenbankservern

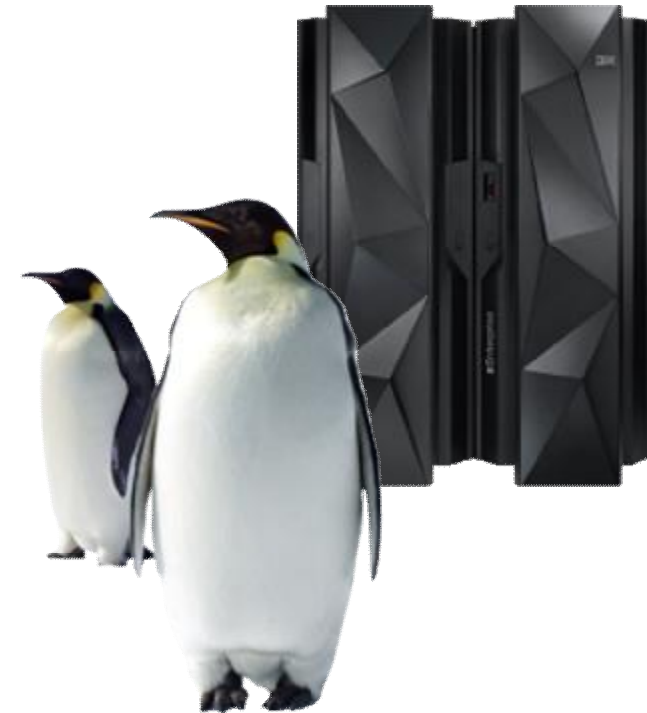
§ Warum Konsolidierung

§ Virtualisierung – die Voraussetzung für eine optimierte und konsolidierte Systemumgebung

§ Praktische Erfahrungen – Kundenbeispiel

- Kundenprofil
- Erfahrungen mit Tuning-Maßnahmen

§ Empfehlungen und „Best Practices“



Nationwide Insurance

- § Durch Konsolidierung in die virtualisierte Infrastruktur, hat Nationwide hunderte von Einzel-Servern ersetzt, die mit mehreren unterschiedlichen Betriebssystemen liefen.
- § ***\$15M Kosteneinsparungen*** in drei Jahren
- § 85% – 90% Serverausnutzung
- § 80% geringere Betriebskosten



Nationwide[™]
Insurar

Beispiel für Kostenreduktion mit Linux auf zEnterprise 114

TCA Analyse (Total Cost of Acquisition):

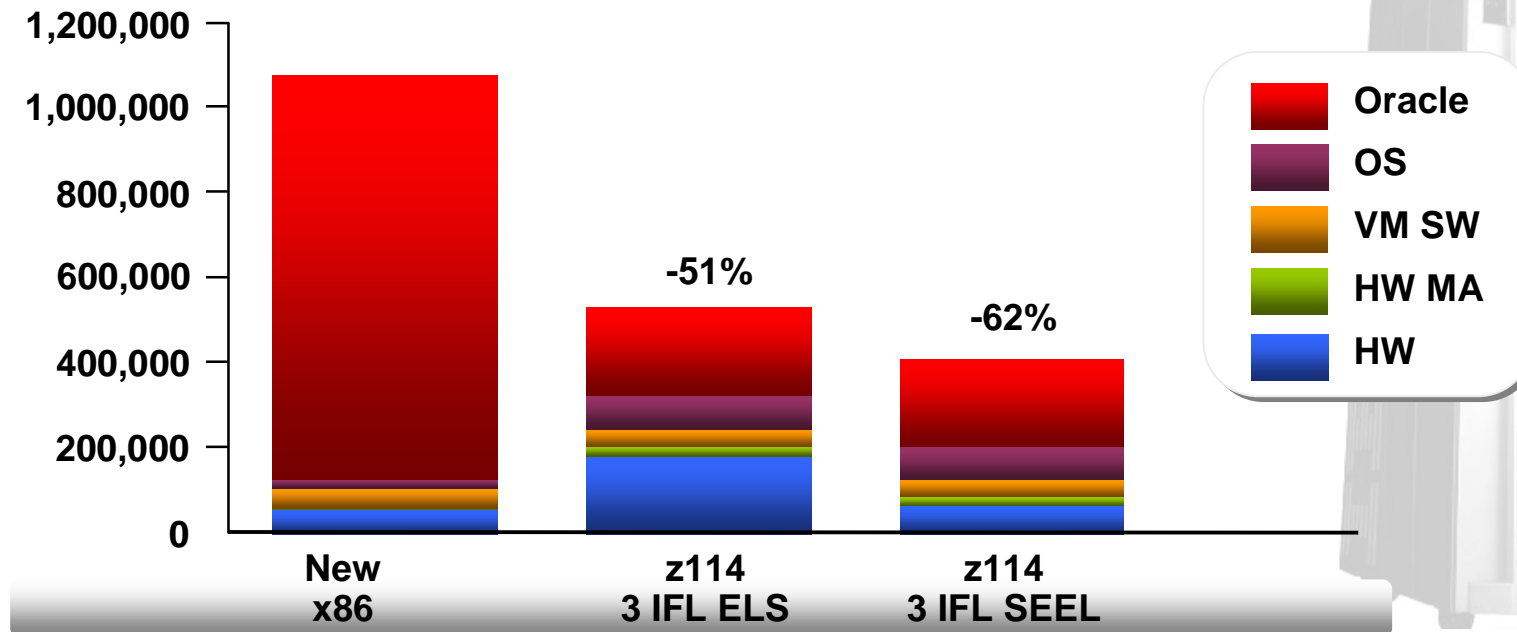
Konsolidierung von 40 Oracle-Servercores auf 3 Linux-Cores auf z114

Geringere Anschaffungskosten für Hardware und Software gegenüber verteilten Servern

– bis zu 51% weniger als Nehalem² mit neuem Rechner (Enterprise Linux Server (ELS)) –

– oder bis zu 62% weniger wenn Server erweitert wird (Solution Edition for Enterprise Linux (SEEL)) –

Plus, zusätzliche Einsparungen bei DR, Stellfläche, Stromverbrauch, Kühlung und Personalkosten



¹ Based on US Enterprise Linux Server pricing. Pricing may vary by country. Model configuration included 10 IFL cores running a mixed workload averaging 31 virtual machines per core with varying degrees of activity. Includes zEnterprise hardware and z/VM virtualization software. Does not include Linux OS or middleware software.
² Distributed server comparison is based on IBM cost modeling of Linux on zEnterprise vs. alternative distributed servers. Given there are multiple factors in this analysis such as utilization rates, application type, local pricing, etc., savings may vary by user.

Reduzierte Softwarekosten durch Konsolidierung

Beispiel: Oracle Datenbank

- § Lizenzkosten und jährliche Software Update Lizenz & Support basieren auf der Anzahl der Prozessorkerne
- § Ein "Processor Core Factor" wird für unterschiedliche Technologien angewandt



<http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf>

Software Update Lizenz & Support (jährlich) ist typischerweise 22% der Prozessorlizenz (Einmalzahlung)

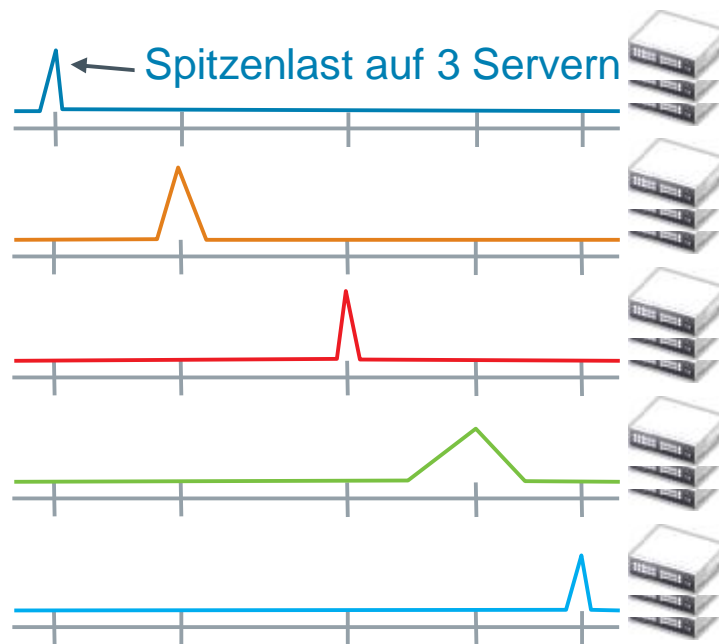


Vendor and Processor	Core Processor Licensing Factor
Intel Xeon Series 56XX, Series 65XX, Series 75XX, Series E7-28XX, Series E7-48XX, Series E7-88XX, Series E5-24XX, Series E5-26XX, Series E5-46XX, Series E5-16XX, Series E3-12XX or earlier Multicore chips	0.5
Intel Itanium Series 93XX (For servers purchased on or after Dec 1st, 2010)	1.0
IBM POWER6	1.0
IBM POWER7	1.0
IBM System z (z10 and earlier)	1.0
All Other Multicore chips	1.0

<http://www.oracle.com/us/corporate/contracts/processor-core-factor-table-070634.pdf>

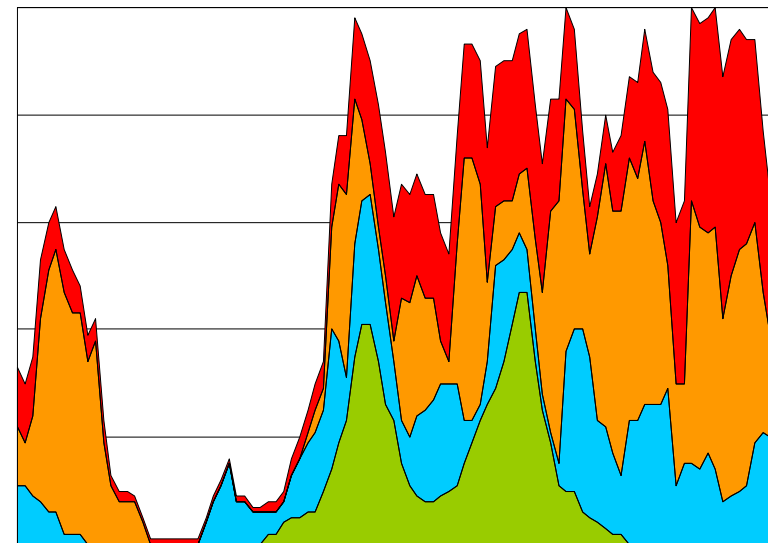
Konsolidierung auf großen High-End Servern

Auslastung auf x86 Systemen



Laut einer Gartner-Studie beträgt die durchschnittliche Server-CPU-Auslastung in Rechenzentren ohne Virtualisierung lediglich 15%.

Auslastung auf großen IBM High End Servern



IBM High End Server:

bis zu 100% CPU-Auslastung

- Hochvirtualisierte und gemeinsam genutzte Ressourcen
- Weniger Server, weniger Stromverbrauch, Kühlung und Administration
- Kostenoptimierte Nutzung der Software

Hohe Core-to-Core Rate für Konsolidierungen von verteilten IT-Umgebungen auf Enterprise Linux Server

Kundenbeispiele mit realer Arbeitslast!

Industry	Distributed Cores	IBM Enterprise Linux Server [™] Cores	Core-to-Core Ratio*
Public	292	5	58 to 1
Banking	111	4	27 to 1
Finance	442	16	27 to 1
Banking	131	5	26 to 1
Insurance	350	15	23 to 1
Insurance	500+	22	22 to 1
Banking	63	3	21 to 1
Finance	854	53	16 to 1
Health care	144	14	10 to 1
Transportation	84	9	9 to 1
Insurance	7	1	7 to 1

* Client results will vary based on each specific customer environment including types of workloads, utilization levels, target consolidation hardware, and other implementation requirements.

Sparda Datenverarbeitung eG nutzt IBM zEnterprise



“Oracle has been consolidated on this platform we are using right now only Oracle on the z196 platform,”

^ Bernd Bohne, Sparda-Datenverarbeitung e.G., Manager,
Central Systems

ibm.com/systems/z/resources/sparda_bank_video.html

“Over the years, the mainframe transformed from traditional workloads, quite simple, to a universal platform for new workloads as well.

And we see a lot of new applications that are coming to this platform. Especially for Linux, it's perfect.

The z/Enterprise platform is perfect for consolidating Linux workloads because of the high I/O bandwidth, business continuity with capacity backup features.”

Diskussionspunkte

Konsolidierung von Oracle Datenbankservern

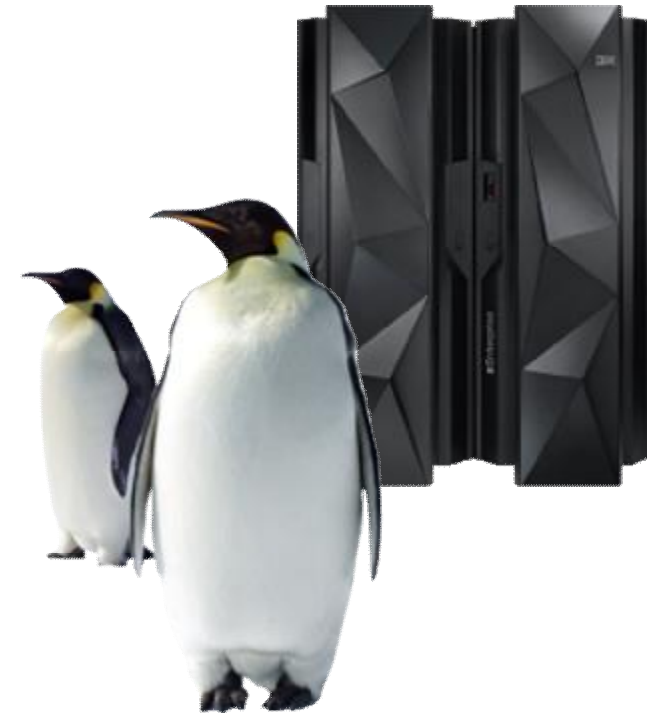
§ Warum Konsolidierung

§ **Virtualisierung – die Voraussetzung für eine optimierte und konsolidierte Systemumgebung**

§ Praktische Erfahrungen – Kundenbeispiel

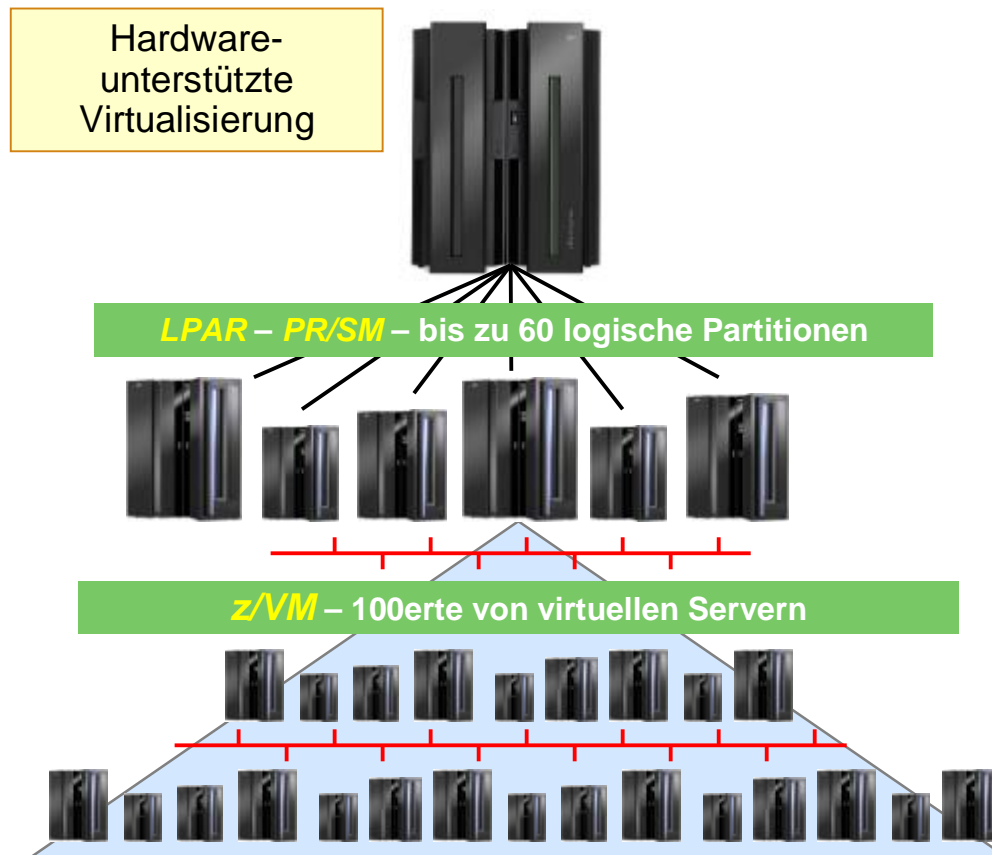
- Kundenprofil
- Erfahrungen mit Tuning-Maßnahmen

§ Empfehlungen und „Best Practices“



System z – Extreme Virtualisierung

Integrierte und gemeinsam genutzte Architektur (Shared Everything)

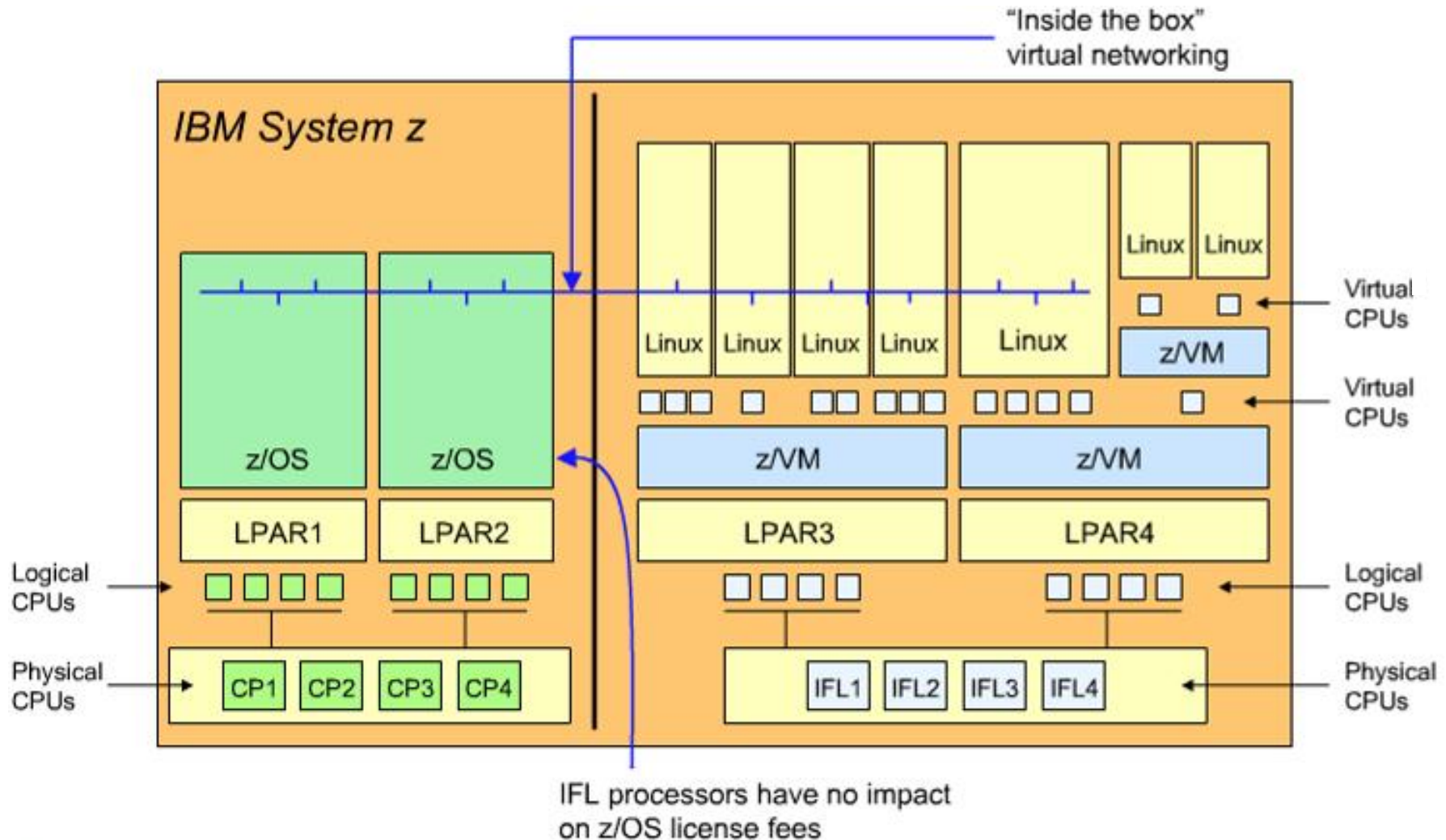


System z

- Inbetriebnahme von virtuellen Servern in Sekunden
- Hohe Granularität der Ressourcennutzung (<1%)
- Aufrüstung der physischen Ressourcen ohne Systemstillstand
- Skalierbarkeit bis zu mehreren 1000 virtuellen Servern
- Mehr mit weniger: mehr virtuelle Server pro Kern, gemeinsame Nutzung der physischen Ressourcen
- Extensives Life-Cycle Management
- HW-unterstützte Isolation, hohe Sicherheit (EAL5 oder EAL4+ zertifiziert)

Extreme Virtualisierung mit z/VM

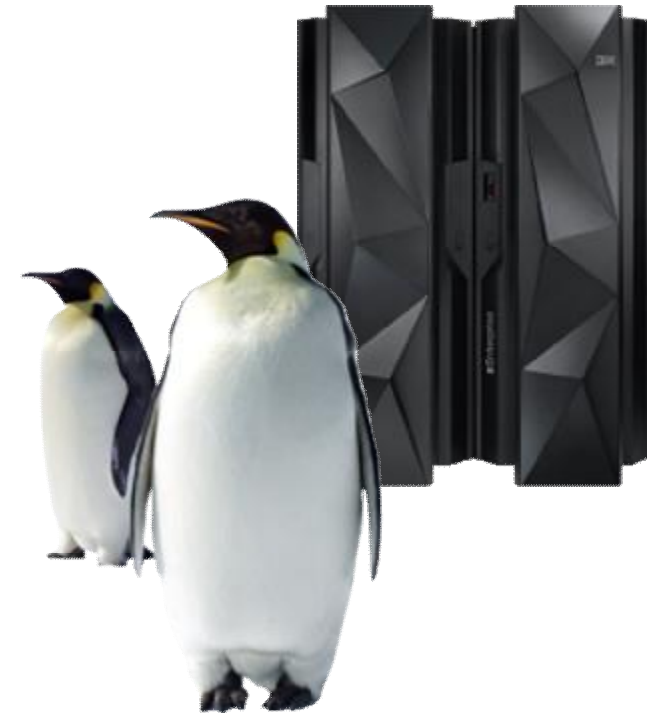
z/VM bietet höchste Skalierbarkeit für eine virtuelle Serverumgebung durch die Kombination von virtuellen und realen Ressourcen für jede virtuelle Maschine



Diskussionspunkte

Konsolidierung von Oracle Datenbankservern

- § Warum Konsolidierung
- § Virtualisierung – die Voraussetzung für eine optimierte und konsolidierte Systemumgebung
- § **Praktische Erfahrungen – Kundenbeispiel**
 - **Kundenprofil**
 - **Erfahrungen mit Tuning-Maßnahmen**
- § Empfehlungen und „Best Practices“



Ein reales Kundenbeispiel

§ Großes Oracle DB Konsolidierungsprojekt

- Oracle 10gR2 Datenbanken (einschließlich einiger 11gR2 Datenbanken)

§ Konsolidierung von x86 (HP ProLiant Blade Server) nach IBM zEnterprise 196

- 16 Prozessorkerne (IFL = Integrated Facility for Linux)
- z/VM V6.1
- RHEL 5.6

§ Migration der einzelnen Datenbanken über einen längeren Zeitraum

- Migrations-Services von IBM Migration Services (“Migration Factory”)

Problembeschreibung:

§ Der Kunde meldete Anwendungs-Laufzeitprobleme auf 3 von etwa 50 Datenbanken

- Business Analyse Anwendung ‘A’: Laufzeit zu lang
- Business Analyse Anwendung ‘B’: Laufzeit zu lang
- Anwendung ‘C’: zunehmende Anzahl von Time-outs (Transaktionen, die 1 Minute überschreiten)

Ein typisches Performance-Problem

Performance-Probleme:

§ Lange Laufzeiten der monatlichen Auswertungs- oder Berichtsprogramme

§ Anwenderteam sagt, dass keine Änderungen an den Anwendungen gemacht wurden

Aber.....

§ Starkes Wachstum der Datenbankgröße

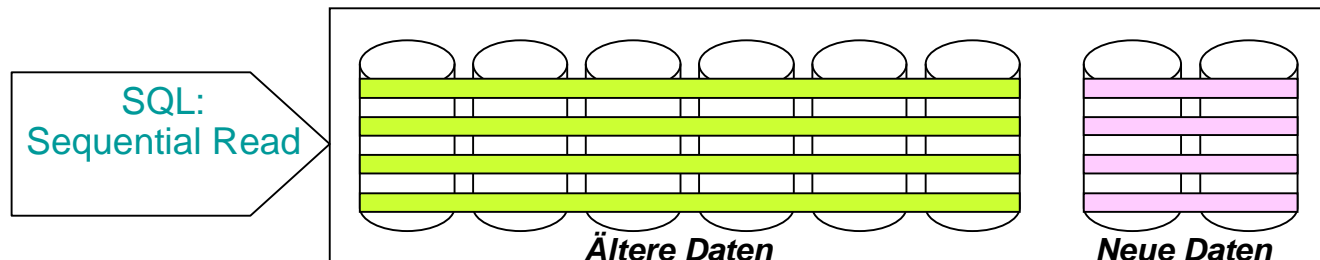
– ca. 12% in 3 Monaten

• April +45 GB, Mai +27 GB, Juni +32 GB

– Vergrößerung des zugeordneten Plattenspeichers - 2 weitere Platten (ca. 360 GB)

§ Späteres Hinzufügen weiterer Plattenspeicher hat Einfluss auf die Datenverteilung (Striping)

– Neuere Daten sind nur über 2 Platten verteilt (2 Platten statt 6)



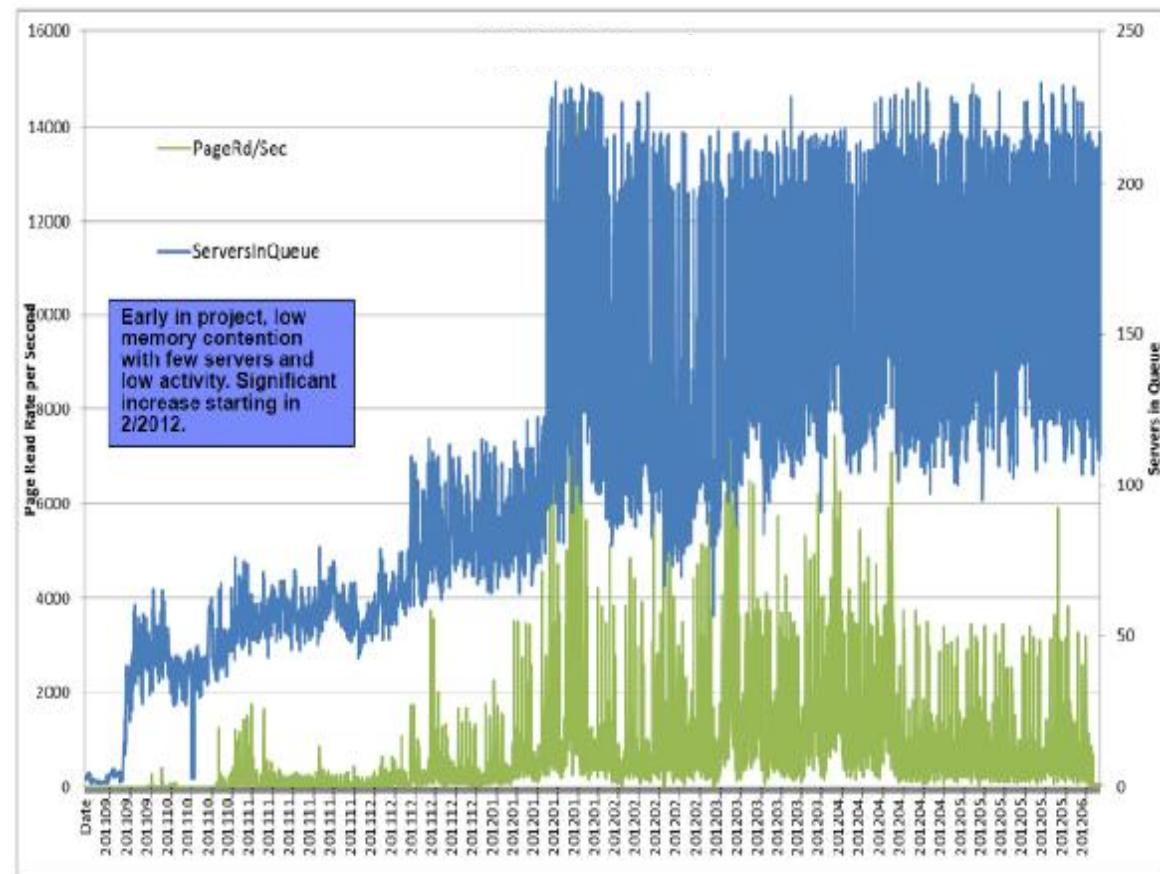
Performance-Verschlechterung über die Zeit

Problem:

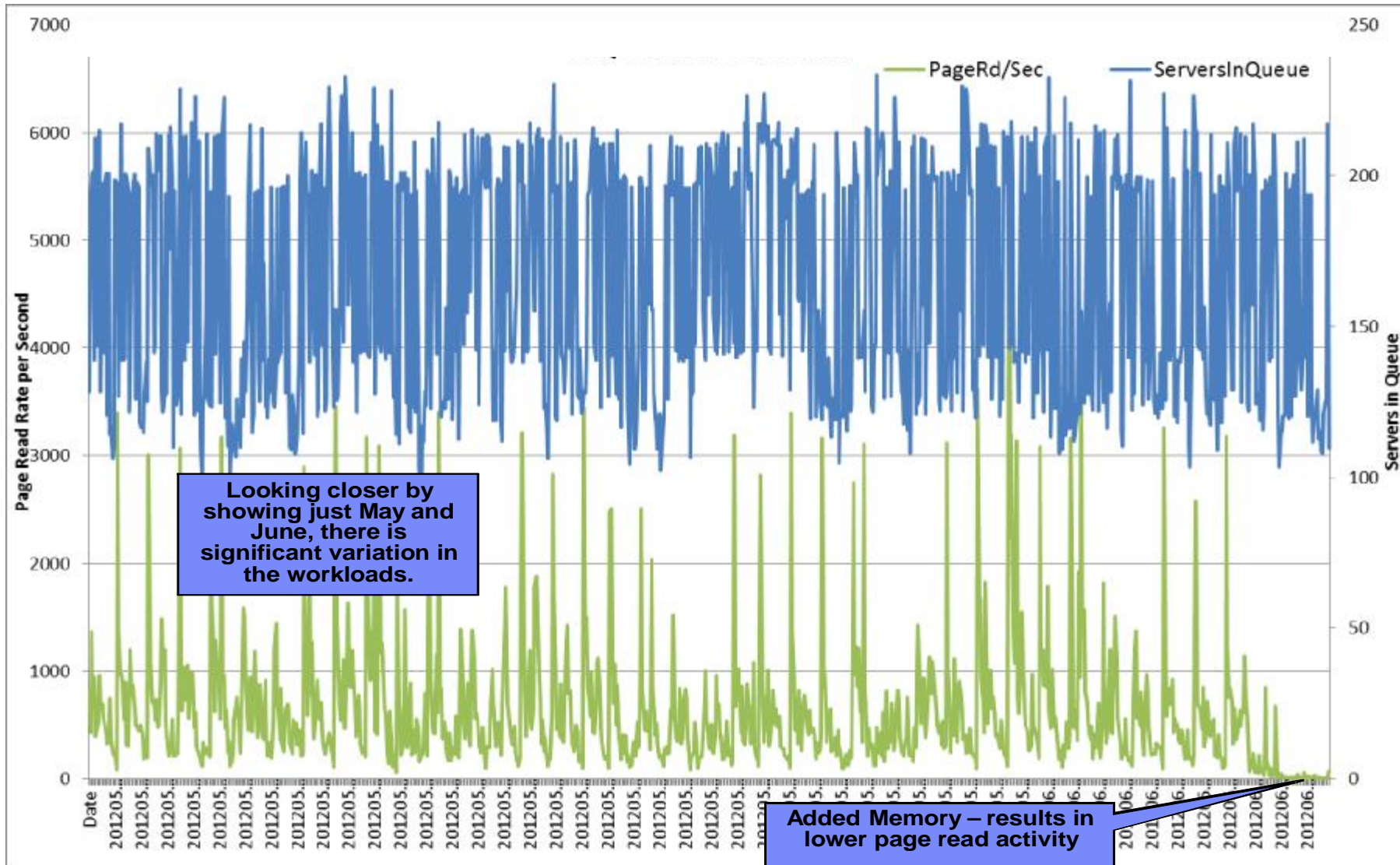
§ Der Durchsatz einiger Server bzw. Datenbankanwendungen verschlechterte sich mit der erhöhten Systemauslastung über die Zeit

Ursache:

- § Die zusätzlichen Server und zunehmende Aktivität führten zu Speicherkonflikten (memory over-commitment)
- § Speicherplatzmangel führte zu hohen Paging-Raten (Auslagerung auf Plattenspeicher) und internem Systemmanagement-Overhead (Wettbewerb um Speicher zwischen den Servern)



Memory Over-commitment Veränderung



Parallelität im Plattenzugriff

§ Plattenspeicher ist direkt den Gästen zugeordnet

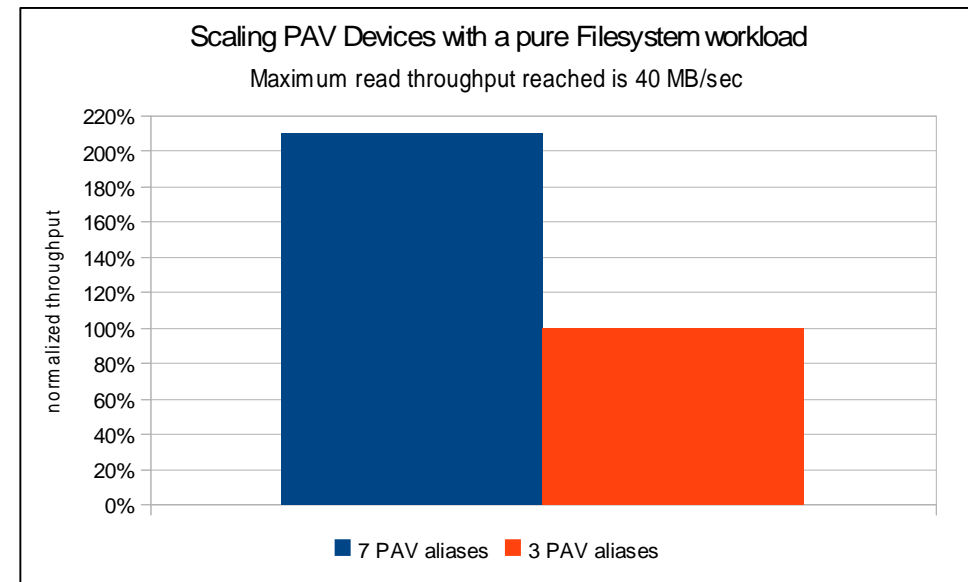
- Dies ist die Empfehlung für I/O-intensive Datenbank-Arbeitslast
- Direct I/O ist Voraussetzung für die Nutzung der PAV Funktion (Parallel Access Volume)

§ Die Anzahl der parallelen Pfade (PAV Devices) ist ein kritischer Parameter für den Durchsatz

§ Mit 8 Pfaden kann das System 2x mehr I/O-Operationen durchführen, als mit 4 Pfaden

§ Messungen zeigten, dass der Plattenzugriff bei 8 Pfaden keinen Flaschenhals darstellt

§ Messwerte sind „random I/O“ Zugriffe (nicht „sequential I/O“)



Oracle DB Tuning Aktivitäten – Business Analyse Anwendung ‘A’

§ Aktionen und Resultate:

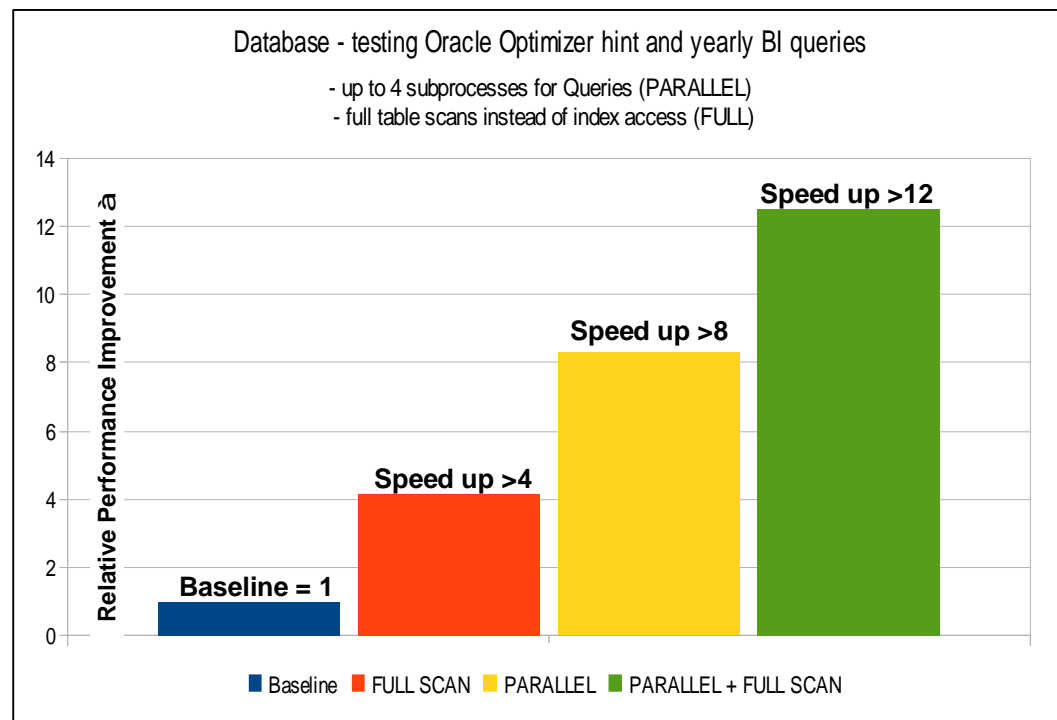
- DB und Anwendung wurden in eine Testumgebung (sandbox) kopiert
 - ∅ Das Problem wurde erfolgreich in der Testumgebung abgebildet
 - ∅ Getestet wurde mit historischen Daten des Vorjahres 2011
- Einsatz von FIO (flexible I/O) Testprogramm um Datenbank-Plattenzugriffe zu simulieren (Stresstest). Ziel: ermitteln der maximal erzielbaren I/O-Bandbreite
 - Anzahl der Pfade (paralleler Zugriff) wurde pro Platte von 4 auf 8 erhöht
 - ∅ Bandbreite erhöhte sich von 4 MB/s auf 8 MB/s
 - Parameter „*rr_min_io*“ wurde von 1000 auf 1 verändert (Linux Vorgabe = 1000)
 - ∅ Bandbreite erhöhte sich von 8 MB/s auf 20 MB/s (im Testfall)
- ∅ **Signifikante Durchsatzverbesserung der Abfragen im Monats- und Jahres-Analyselaufr**

§ Tests mit Oracle Optimizer zeigten weitere dramatische Verbesserungen

Oracle DB Tuning Aktivitäten – Business Analyse Anwendung ‘A’

§ Oracle Optimizer Hints gelten nur für die spezifischen SQL-Anweisungen

- 'FULL' forciert *table scans* statt *index access*
- 'PARALLEL' forciert ein Aufbrechen der Anweisungen, sodass eine parallele Ausführung möglich ist
- 'PARALLEL' und 'FULL'



§ Risiko

- Forcieren von *table scan* kann zu einer erheblichen Performanceverschlechterung führen, wenn *index access* die passende Zugriffsmethode wäre.
- Es kann Gründe geben, die eine parallele Ausführung für bestimmte Anweisungen nicht zulässt, wodurch keine Verbesserung erzielt würde.

Oracle DB Tuning Aktivitäten – Business Analyse Anwendung ‘B’

§ Workflow für Datenanalyse

- DB in Testumgebung kopiert
- Anwendung besteht aus
 - 3 Schritten (S, R, und D) mit unterschiedlichen Arbeitsschritten (workflows)
 - Nur die letzten 2 Teile (R und D) sind performance-kritisch

§ **Ausgangslage: 13 Stunden** Laufzeit für Analyse mit Daten eines Jahres

- Ursprüngliche Laufzeit der Anwendung nach Konsolidierung

§ **Test 1** (Laufzeit 07:12:31)

- Tuning der Umgebung (Speicher, Plattenkonfiguration, etc.)
- Verbesserung annähernd Faktor 2

§ **Test 2** (Laufzeit 06:57:57)

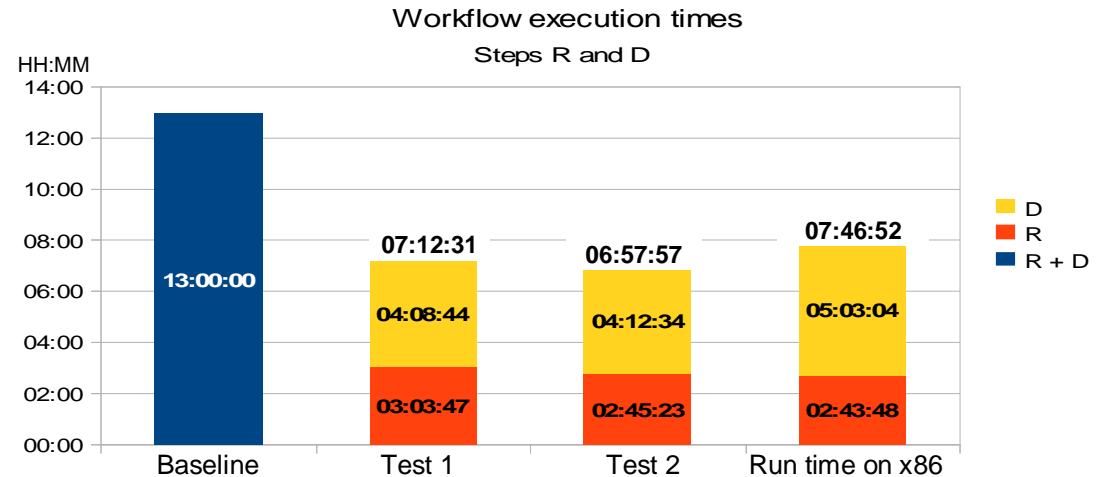
- Verbesserungen von Test 1 und
- Datenbank-spezifisches Tuning (Oracle Parameter)
- Beide Tuning-Schritte zusammen führten zu einer Verbesserung um mehr als Faktor 2 gegenüber der Ausgangslage

Oracle DB Tuning Aktivitäten – Business Analyse Anwendung ‘B’

Änderungen der Parameter:

§ Test 1 (run time 07:12:31)

- Added memory to LPAR
- Enabled 7 PAV devices per DASD device, directly attached to the guest,
- Multipath setup: round robin with `rr_min_io=1`



§ Test 2 (run time 06:57:57)

- Ensure that huge pages are really used → caused a SGA reduction from 8192MB to 7600MB (*better solution would have been to increase the amount of configured huge pages*)
- Profile parameter changes:
 - `db_writer_processes=2` (prior 8),
 - `filesystemio_options=setall` (prior `asynch`),
 - `parallel_degree_policy=auto` (prior `manual`),
 - `pga_aggregate_target=3700M` (prior 3,221,225,472)
- Added parameters:
 - `log_buffers=104,857,600`
- Removed parameters:
 - `disk_asynch_io`,
 - `log_checkpoint_timeout`,
 - `optimizer_index_caching`,
 - `optimizer_index_cost_adj`,
 - `shared_pool_size`

Oracle DB Tuning Aktivitäten – Anwendung ‘C’

Oracle DB für Windows Abwendungs-Server – transaktionale Arbeitslast

- § **Kritisches Limit:**
- Anfragen sollten in weniger als 60 Sekunden beendet sein
 - Maximal 30 Time-outs (>60 sec) werden pro 24 Stunden-Intervall akzeptiert

	Known as Good case	Problem Case	After tuning action part 1	After tuning action part 2
Measurement Duration	24 h	23 h	17.25 h	48 h
Less than 3 Sec	91,79%	88,37%	88,31%	99,97%
3 to 5 Sec	2,74%	3,35%	3,69%	0,02%
5 to 10 Sec	2,74%	3,50%	3,20%	0,01%
10 to 60 Sec	2,58%	4,48%	4,27%	0,00%
More than 60 Sec	0,16%	0,30%	0,53%	0,00%
More than 60 Sec	13 requests	29 requests	24 requests	0 requests

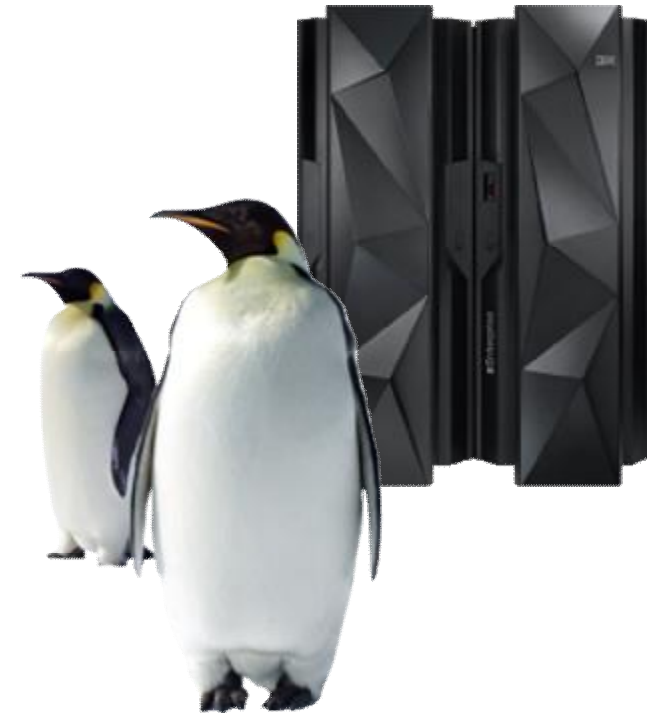
- § **Tuning Aktion Teil 1:**
- Plattenpfade von 4 auf 8 erhöht
 - `rr_min_io = 1`
 - Inaktive Server abgeschaltet (Reduzierung des Drucks auf realen Speicher)
 - Analyse zeigte Korrelation mit Swapping Aktivitäten – virtueller Speicher des Linux Gastes wurde um 2 GB erhöht und direct I/O aktiviert
 - Ø **Monitoring zeigte verbesserte Werte, aber noch immer Time-outs**

- § **Tuning Aktion Teil 2:**
- Erhöhung der vCPUs von 2 auf 4, SGA um 2 GB vergrößert
 - Ø **Dramatische Verbesserung – keine Time-outs**
 - Ø **Resultate wurden auch über längere Laufzeit bestätigt**

Diskussionspunkte

Konsolidierung von Oracle Datenbankservern

- § Warum Konsolidierung
- § Virtualisierung – die Voraussetzung für eine optimierte und konsolidierte Systemumgebung
- § Praktische Erfahrungen – Kundenbeispiel
 - Kundenprofil
 - Erfahrungen mit Tuning-Maßnahmen
- § Empfehlungen und „Best Practices“



Generelle Empfehlungen – Monitoring

Etablieren eines permanenten Monitorings

§ z/VM Performance Toolkit

- Die Hypervisor-Sicht
- **Stellen Sie sicher, dass das Verhältnis von virtuellem zu realem Speicher in einem angemessenen Verhältnis steht**
- Test und Entwicklung erlauben höheres Speicher Over-commitment als Produktion

- Indikatoren:

- z/VM Paging activity

- Report 'User Paging Activity and Storage Utilization' (UPAGE, FCX113)

- Columns: 'X>DS' paging to DASD, critical: Reads paging from DASD

- z/VM Guest Waits

- Report 'Wait State Analysis by User' (USTAT,FCX114)

- Especially columns %PGW, %PGA, and %CFW

- z/VM CPU load

- Report 'System Performance Summary by Time' (SYSSUMLG, FCX225)

- Report 'General CPU Load and User Transactions' (CPU, FCX100)

§ Linux sadc/sar

- Die Linux Server-Sicht

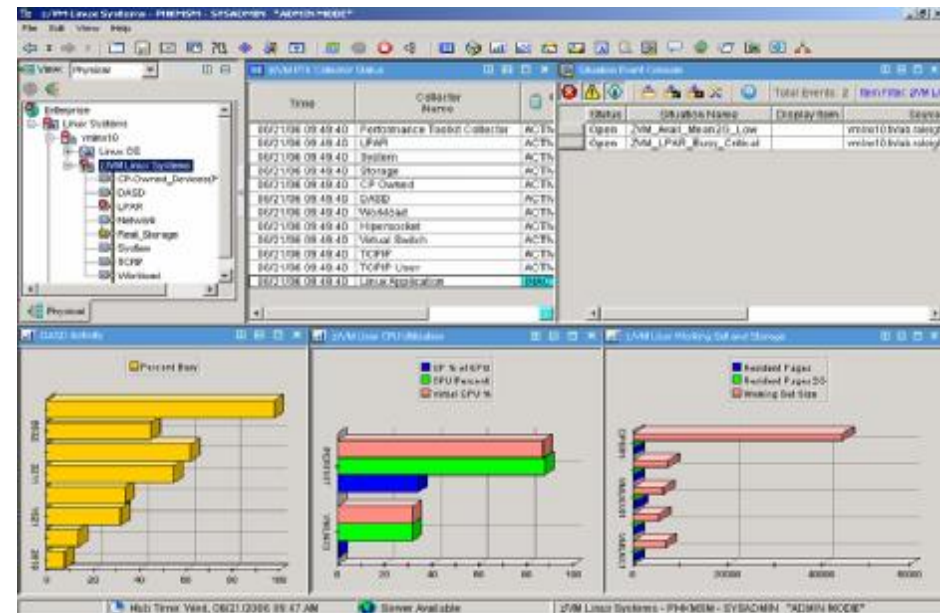
System Management und Monitoring

§ Pro-aktives System Management

– Potentielle Probleme und Flaschenhälse erkennen, bevor die Benutzer klagen!

§ Kapazitätsplanung

§ Kosten weiterbelasten – Charge back



Monitor-Produkte:

§ IBM Tivoli OMEGAMON® XE on z/VM® and Linux

– Tivoli Composite Application Manager (ITCAM) for Applications – Oracle Agent

§ Velocity Software - ZVPS For Managing z/Linux and z/VM Performance

Generelle Empfehlungen - Linux

Speicherbedarf:

§ Virtuelle Speichergröße für Linux sollte nicht zu groß bemessen werden, weil -

- Linux nutzt überschüssigen Speicher für I/O-Puffer und File System Cache.
- In einer virtualisierten Umgebung unter z/VM führen übergroße (oversized) Gäste zu unnötigem Stress auf das VM Paging Subsystem.
- Realer Speicher ist eine gemeinsam genutzte (shared) Ressource. Caching von Seiten im Linux Gast verringert den Speicher, der für andere Gäste zur Verfügung steht.
- Größere virtuelle Speicherzuordnungen benötigen mehr Speicher für Address Space Management im Kernel.

§ Potentiell für alle Datenbanksysteme *vm.swapiness* auf 0 setzen (sysctl.conf)

- Definiert eine Präferenz den Page Cache zu nutzen, statt Swap durch Linux

Beispiel für Speicherbedarfsplanung

§ Standard Speicherbedarfsplanung (virtueller Linux-Speichergröße) = Summe von:

- Speicher für Linux Kernel: 512 MB
- Speicher für Oracle SGA: laut DBA Abschätzung
- Speicher für Oracle PGA: laut DBA Abschätzung
- Speicher für Oracle ASM: 256 MB bis 512 MB (falls ASM verwendet wird)
- Speicher für zusätzliche Agenten, wie OEM, Tivoli etc., nach Bedarf der Anwendung
- Linux Overhead: 5 % des gesamten Speichers für den Gast

Startgröße = SGA + PGA + 0.5GB für Linux + ASM (falls verwendet)

§ Speicher Over-commitment (Verhältnis von virtuellem zu realem Speicher)

- Kein oder geringes Speicher Over-commitment für kritische Produktions-Datenbanken
- Test und Development Gäste können von der z/VM Speicher Over-commitment-Funktion profitieren

Generelle Empfehlungen – Oracle Parameter

§ **Empfehlung:** Parameter *filesystemio_options=setall*

- Vermeidet das Zwischenspeichern (caching) von I/O im Datenbank Puffer Cache von Oracle und zusätzlich im File System Cache von Linux und nutzt “direct I/O”.

§ **Definition des Log-Puffers:** *log_buffer = 104857600* oder größer

§ **Log Setup:** Plazieren der Logs auf separaten Platten

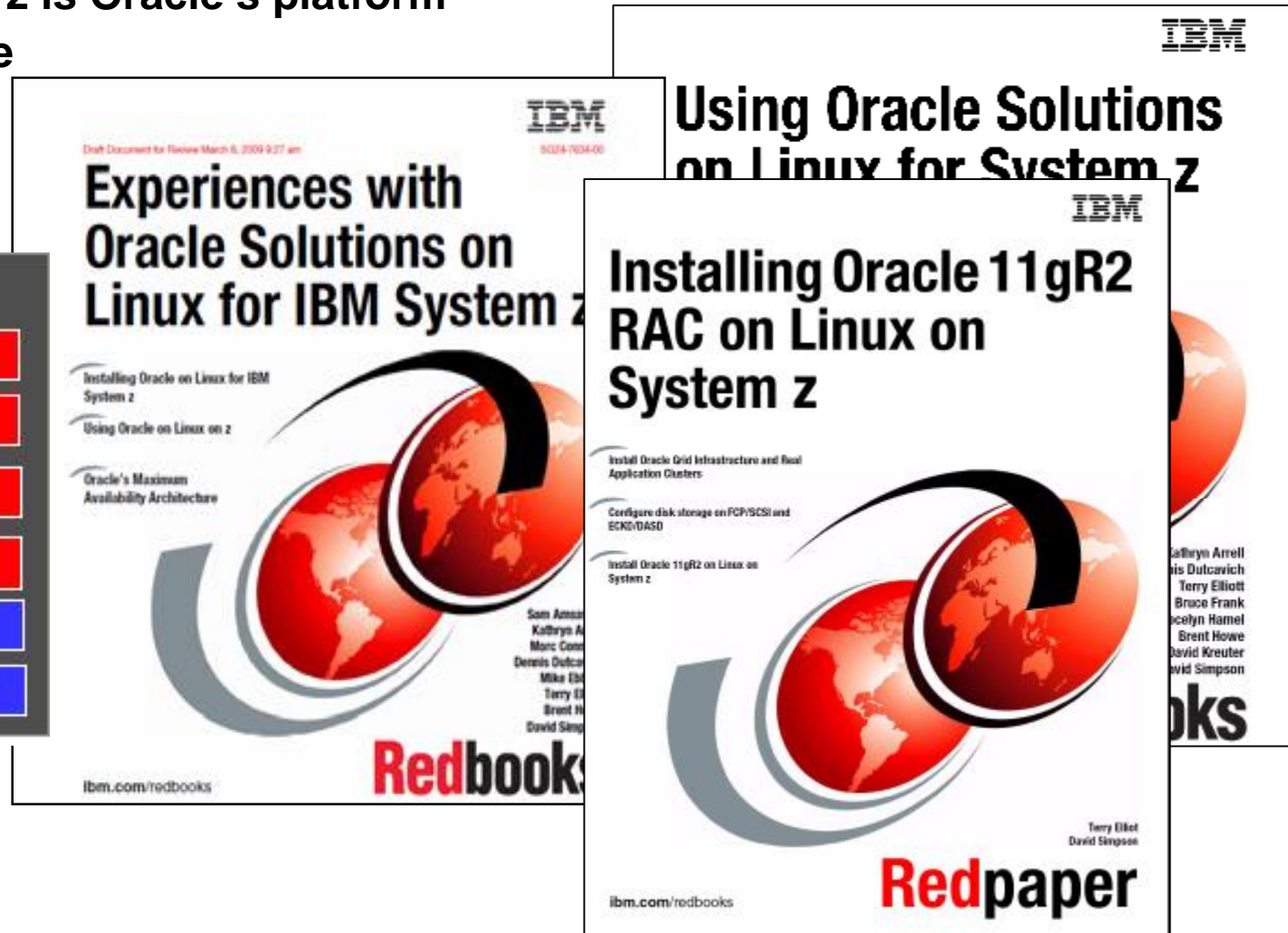
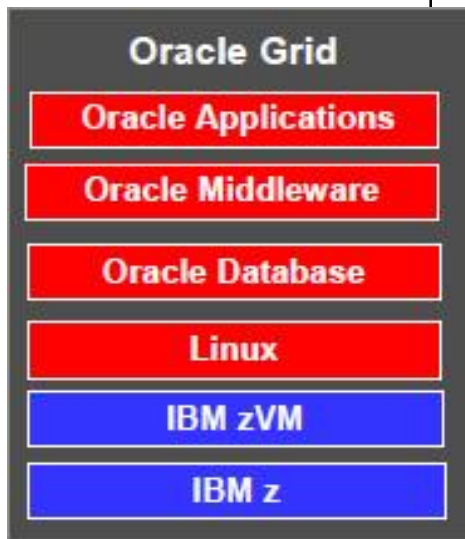
- Größere Log-Files sind besser, da Umschalten der Logs seltener vorkommt

§ Existierende **Optimizer Hints** überprüfen!

- Vorsicht mit der Definition von **Optimizer Parametern** (*optimizer_...*) als globale Parameter, da dies nur für einige Anwendungen von Vorteil sein könnte.
 - Optimizer Hints in den SQL-Befehlen sind wahrscheinlich besser, da nur für spezifisch SELECT Befehle gültig.
- Erzielten starke Verbesserungen mit FULL(<table name>) und PARALLEL(<table name>, <number of CPUs>) für Business Analyse Abfragen
- Ein typischer Parallelitätsgrad ist <amount of vCPUs> oder <amount of vCPUs + 1>, das obere Limit sollte nicht größer als 2 mal die Anzahl der virtuellen CPUs des Gastes sein.
- Für Oracle 11g kann *parallel_degree_policy=AUTO* statt explizierter Optimizer Hints besser sein, um Oracle die Wahl des Parallelisierungsgrades zu überlassen.

Oracle and Linux on System z – IBM & Oracle working together

Linux on System z is Oracle's platform for the mainframe



Warum Oracle auf System z

- § Hohe Anforderungen an Verfügbarkeit
- § Disaster Recovery Anforderungen
- § Wachsende Performance-Anforderungen
- § Gesamtkosten (TCO) gegenüber Anschaffungskosten (TCA)
- § Umweltverträglichkeit (Strom, Kühlung)
- § Extreme Virtualisierung mit bis zu 100% Auslastung
- § System z hat die höchst Sicherheitseinstufung und -klassifizierung eines kommerziellen Servers



Questions?



Siegfried Langer
Business Development Manager
z/VSE & Linux on System z



IBM Deutschland Research
& Development GmbH
Schönaicher Strasse 220
71032 Böblingen, Germany

Phone: +49 7031 - 16 4228

Siegfried.Langer@de.ibm.com

Notices & Trademarks

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to: IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

This presentation contains trade-marked IBM products and technologies. Refer to the following Web site:

<http://www.ibm.com/legal/copytrade.shtml>