

SOUG Meeting: Exadata zum Anfassen

# Architektur der Exadata

25. November 2010

Manfred Drozd



in&out

# Inhalt



- 
- 1 Serienproduktion anstatt manueller Fertigung
  - 2 Exadata Database Server
  - 3 Exadata Storage Server
  - 4 Exadata Flash Technologie
  - 5 Arbeiten mit der Exadata
  - 6 Zusammenfassung
-

# Über die Firma In&Out AG



- Schweizer Consulting Unternehmen
- 1993 gegründet, 35 Mitarbeiter
- 2 Geschäftsbereiche
  - IT Security
  - IT Efficiency: Strategie, Architektur, Evaluation & Performance Benchmarking von IT Plattformen
- **Herstellerunabhängig**
- Benchmark Tools: CPUbench, Iogen, OraBench

# Manuelles Engineering

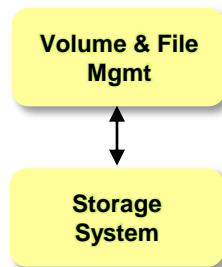


**Storage  
System**

***Komplexität von Datenbank Plattformen***

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.



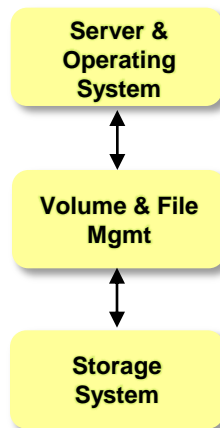
***Komplexität von Datenbank Plattformen***

## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS, VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async, direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.



***Komplexität von Datenbank Plattformen***

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen, (x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading, Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme, > 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

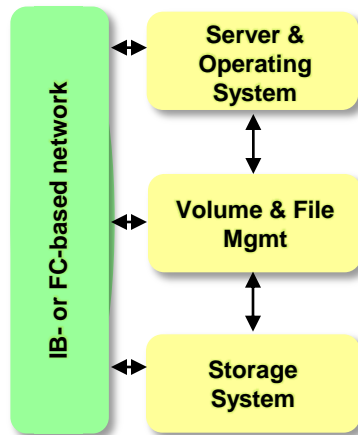
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS, VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async, direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.

# Manuelles Engineering



**Komplexität von Datenbank Plattformen**

## **FC-basierte oder IB-basierte Netzwerke**

Bandbreite , Latenzzeit bei Remote Storage Spiegelung (sync, async) wegen Switches, Hubs und Distanz.

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen, (x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading, Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme, > 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

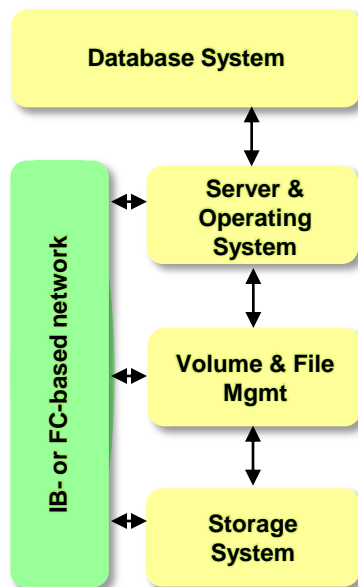
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS, VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async, direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.

# Manuelles Engineering



**Komplexität von Datenbank Plattformen**

## **Oracle Database**

Verschiedene Versionen, Patches und Optionen,  
> 100 Konfigurationsparameter

## **FC-basierte oder IB-basierte Netzwerke**

Bandbreite , Latenzzeit bei Remote Storage Spiegelung (sync, async)  
wegen Switches, Hubs und Distanz.

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen,  
(x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading,  
Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme,  
> 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

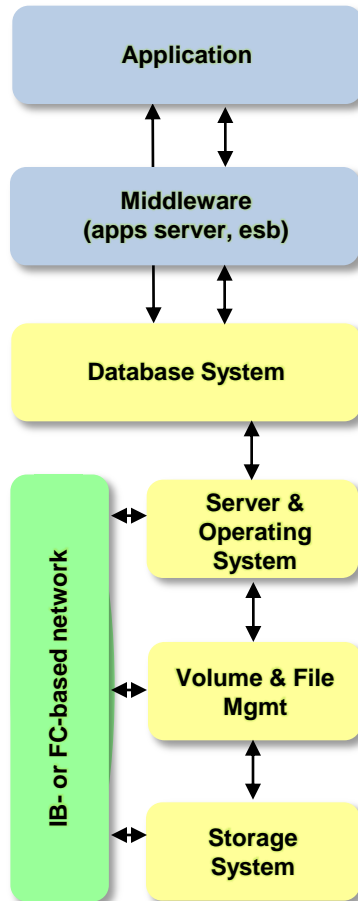
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS,  
VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async,  
direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o  
unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage  
Technologien: Anzahl Disks und Geschwindigkeit, RAID Management,  
Cache Management, Interface Technologie, Storage System Optionen  
wie Remote Copy, Hardware Striping und/oder Spiegelung,  
Virtualisierung von Ressourcen.

# Manuelles Engineering



**Komplexität von Datenbank Plattformen**

## **Oracle Database**

Verschiedene Versionen, Patches und Optionen,  
> 100 Konfigurationsparameter

## **FC-basierte oder IB-basierte Netzwerke**

Bandbreite , Latenzzeit bei Remote Storage Spiegelung (sync, async)  
wegen Switches, Hubs und Distanz.

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen,  
(x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading,  
Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme,  
> 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

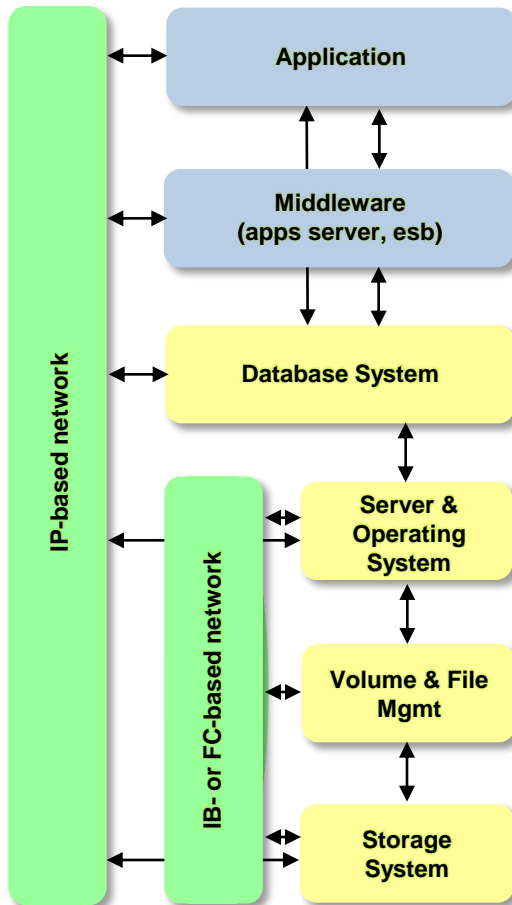
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS,  
VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async,  
direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o  
unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systems, Storage Tiers und Storage  
Technologien: Anzahl Disks und Geschwindigkeit, RAID Management,  
Cache Management, Interface Technologie, Storage System Optionen  
wie Remote Copy, Hardware Striping und/oder Spiegelung,  
Virtualisierung von Ressourcen.

# Manuelles Engineering



**Komplexität von Datenbank Plattformen**

## **IP-basierte Netzwerke**

Durchsatz & Latenzzeit für Remote Database Spiegelung (sync, async) wegen Switches und Protokollen wie SQL\*Net und TCP/IP

## **Oracle Database**

Verschiedene Versionen, Patches und Optionen, > 100 Konfigurationsparameter

## **FC-basierte oder IB-basierte Netzwerke**

Bandbreite, Latenzzeit bei Remote Storage Spiegelung (sync, async) wegen Switches, Hubs und Distanz.

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen, (x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading, Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme, > 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

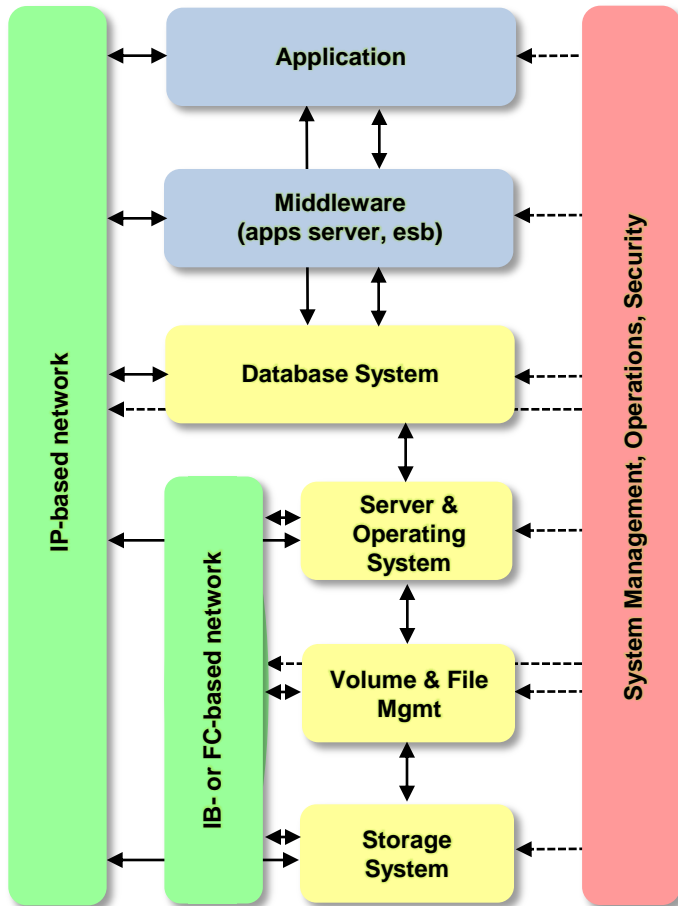
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS, VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async, direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o unit), Software Striping und/oder Spiegelung, Multipathing.

## **Storage System**

Verschiedene Storage Systeme, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.

# Manuelles Engineering



*Komplexität von Datenbank Plattformen*

## **IP-basierte Netzwerke**

Durchsatz & Latenzzeit für Remote Database Spiegelung (sync, async) wegen Switches und Protokollen wie SQL\*Net und TCP/IP

## **Oracle Database**

Verschiedene Versionen, Patches und Optionen, > 100 Konfigurationsparameter.

## **FC-basierte oder IB-basierte Netzwerke**

Bandbreite, Latenzzeit bei Remote Storage Spiegelung (sync, async) wegen Switches, Hubs und Distanz.

## **Server & Betriebssysteme**

Verschiedene Server Systeme, Prozessoren und CPU Architekturen, (x86, IA-64, UltraSparc, SPARC64, Power), #Cores, Multithreading, Hauptspeicherkapazität, Bus Architekturen, Betriebssysteme, > 100 Konfigurationsparameter, Virtualisierung von Ressourcen.

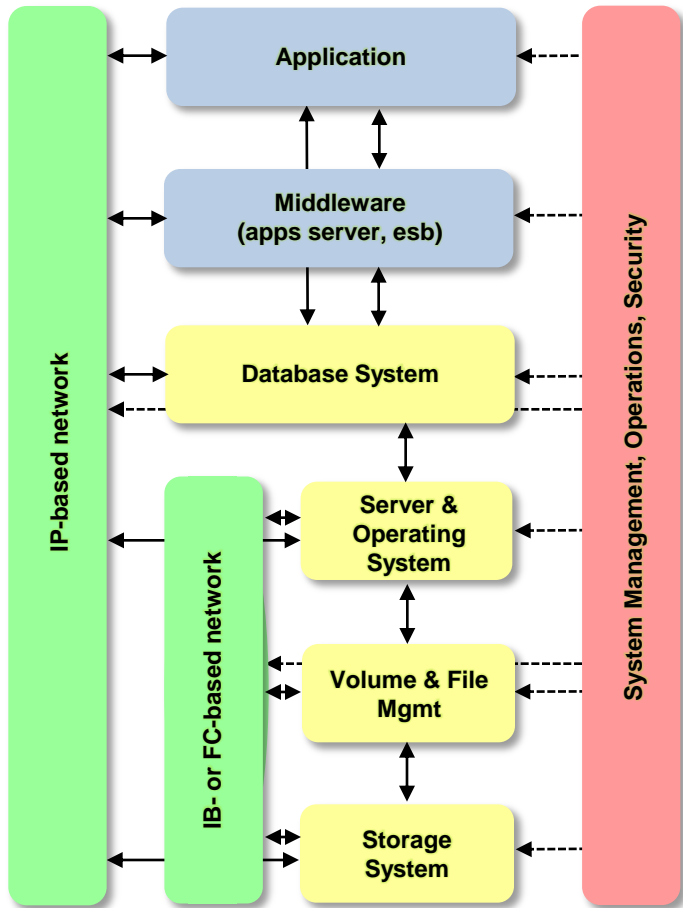
## **Volume & File Management**

Verschiedene Volume Manager (VxVM, ASM) und File Systeme (UFS, VxFS, ext3, JFS, ZFS, raw devices), verschiedene I/O Methoden (async, direct), viele Konfigurationsparameter (#LUNS, queue depth, max i/o unit), Software Striping und/oder Spiegelung, Multipathing.

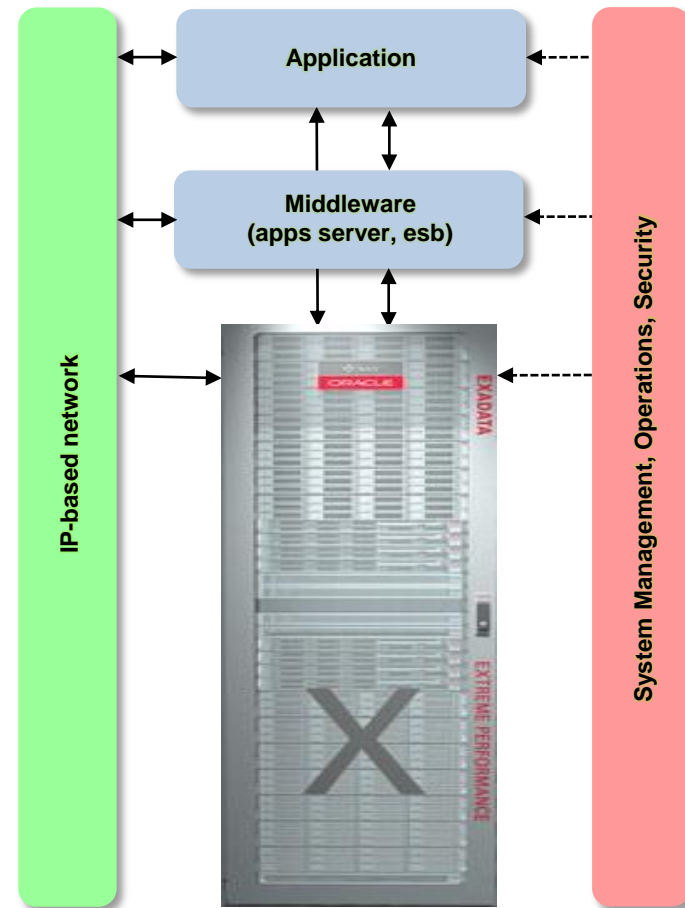
## **Storage System**

Verschiedene Storage Systeme, Storage Tiers und Storage Technologien: Anzahl Disks und Geschwindigkeit, RAID Management, Cache Management, Interface Technologie, Storage System Optionen wie Remote Copy, Hardware Striping und/oder Spiegelung, Virtualisierung von Ressourcen.

# Serienproduktion

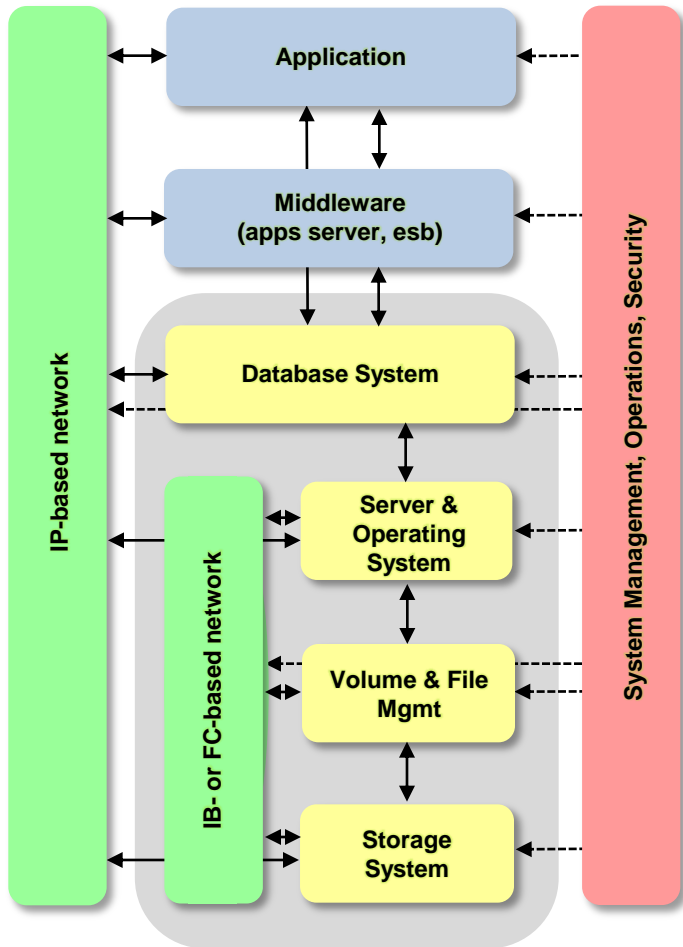


*Best of Breed*

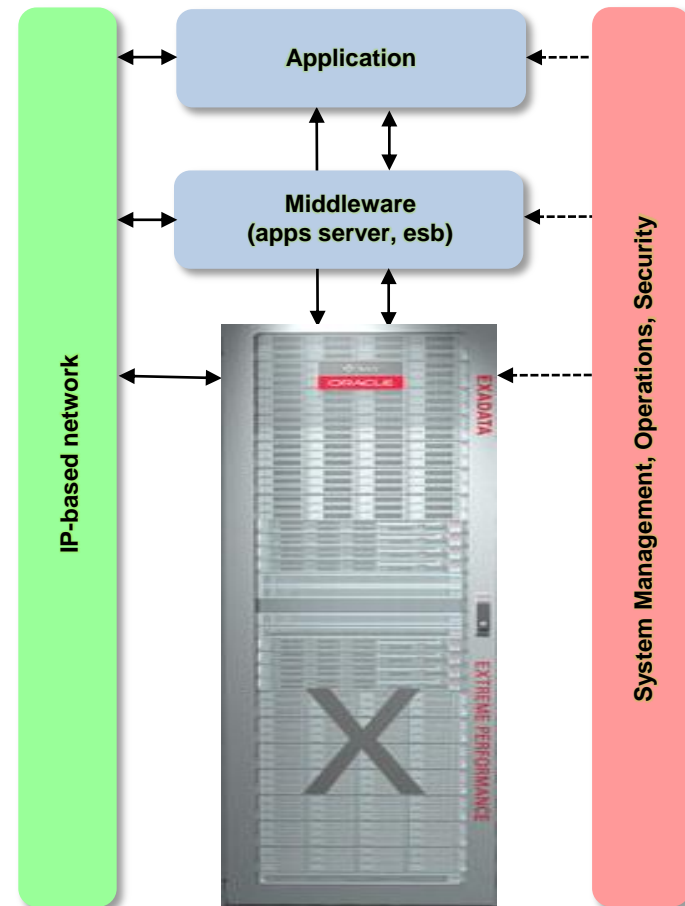


*Oracle Database Machine*

# Serienproduktion



*Best of Breed*



*Oracle Database Server Solution*

# Exadata Database Server



## ■ Standard Server

- Sun Fire X4170
- 2 x Intel Xeon E5540,  
8 Cores, 2.53 GHz
- 72 GByte DDR3 RAM
- 4 x 146 GByte disks
- 10k rpm SAS
- RAID Controller  
512 Mbyte Cache
- Dual-Port Sun QDR  
InfiniBand HCA 40 Gbit/sec

## ■ Kosten

- Listenpreis Database  
Server  
< 20k USD
- Listenpreis 128 GByte  
RAM, 4 GByte DIMM  
< 10k USD

# Exadata Database Server



- Gartner zur Entwicklung des x86 Marktes:

*“By the next decade, the server market will have reached consolidation to one primary technology — x86 — with RISC/Itanium technologies fighting for dwindling market share.”*

*“By 2015, x86-based servers will reach a worldwide shipment share of 97.4%.”*

Quelle: *Impact of the New Generation of x86 on the Server Market*;  
George J. Weiss, Jeffrey Hewitt; Gartner Research June 2010, ID Number: G00201232

# Exadata Storage Server



## ■ Standard Server

- Sun Fire X4275
- 2 x Intel Xeon E5540,  
8 Cores, 2.53 GHz
- 24 GByte DDR3 RAM
- 12 x 600 GByte disks
- 15k rpm SAS
- RAID Controller  
512 MByte Cache
- Dual-Port Sun QDR  
InfiniBand HCA 40 Gbit/sec
- 4 x Sun F20 SmartFlash,  
je 96 GByte

## ■ Kosten

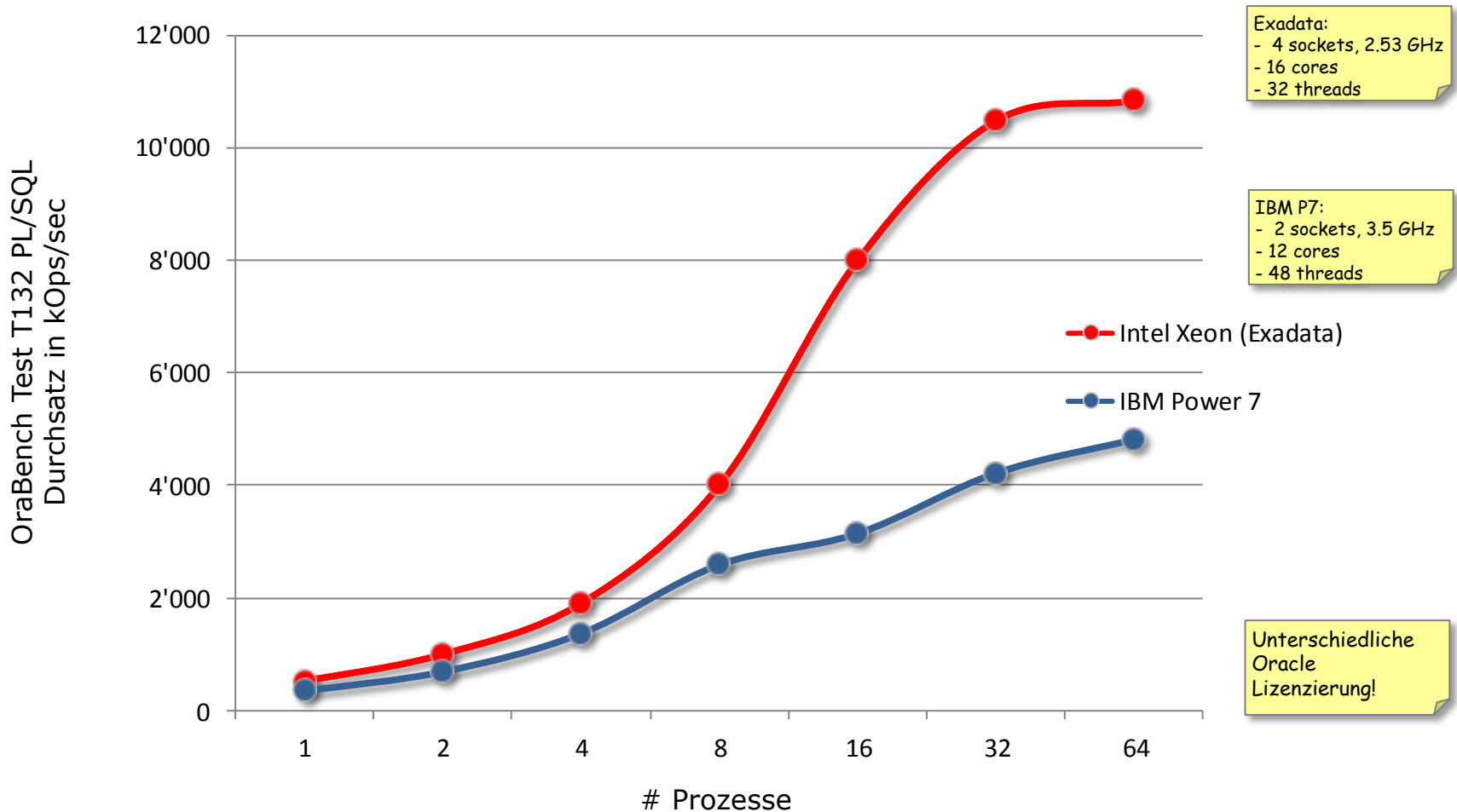
- Listenpreis Storage Server  
Transaktionsorientiert  
< 50k USD
- Listenpreis Storage Server  
Kapazitätsorientiert  
< 40k USD

# Sophisticated Software



- Oracle hat nun Kontrolle über alle Layer und kann Komponenten optimal aufeinander abstimmen
  - Operating System
  - Volume Manager und File System
  - Database System
  - Cache Management (*database server, storage server*)
  - ...
  
- Exadata bietet zusätzliche Funktionalität
  - Offload Funktionen, z.B. *smart scan*
  - Storage Index
  - Neue Komprimierungsverfahren
  - ...

# Exadata Database Server CPU Performance



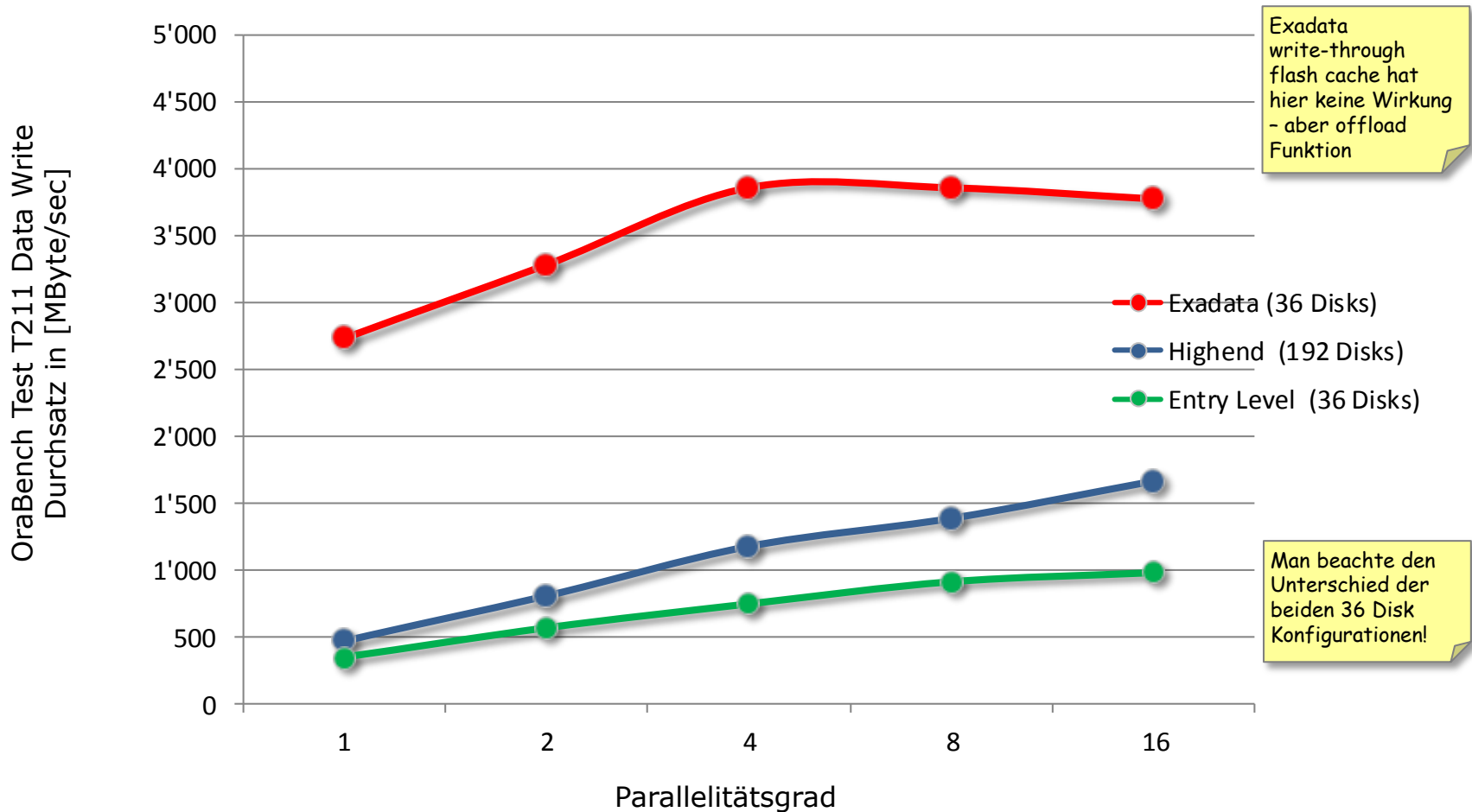
# Exadata Database Server CPU Performance



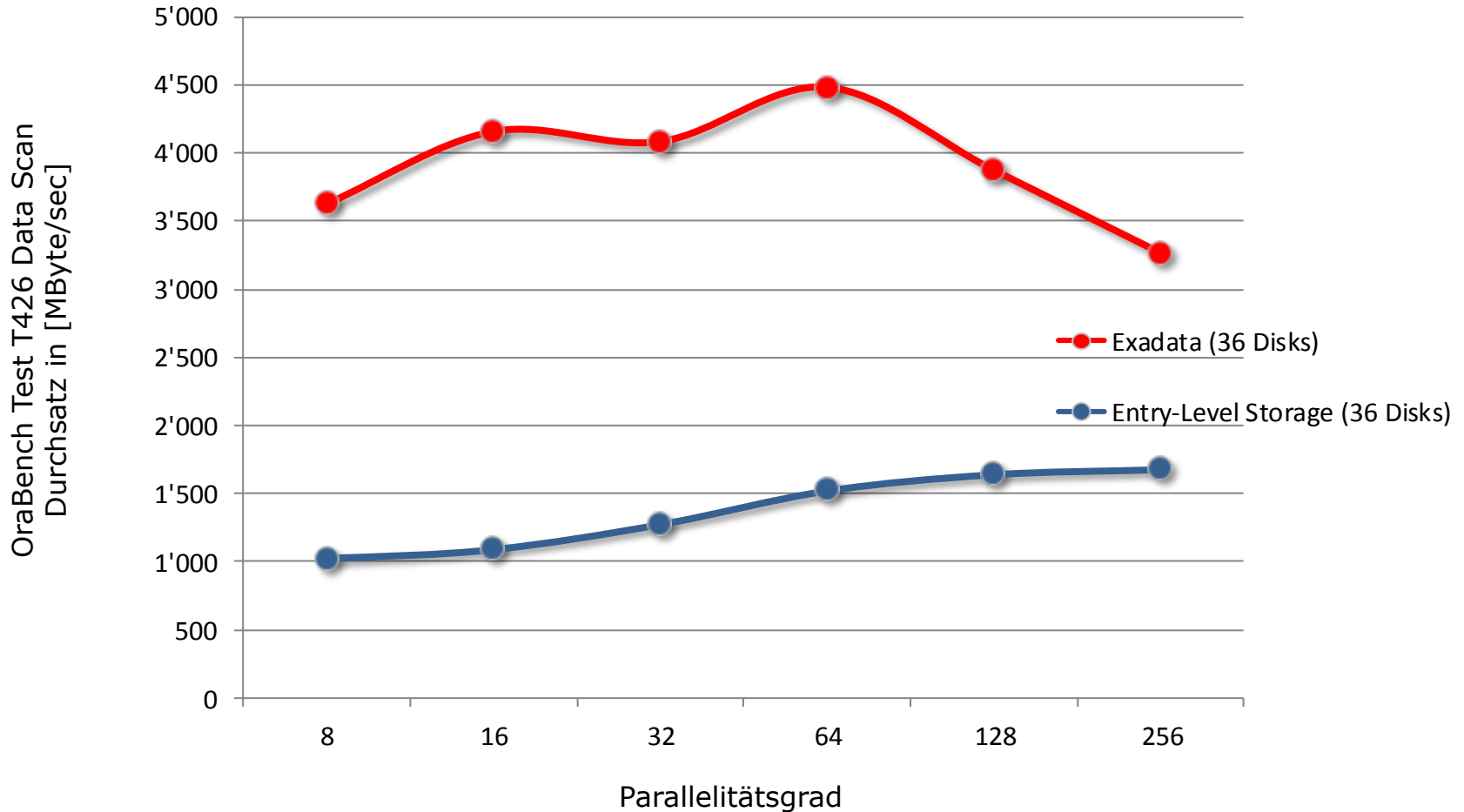
- Datenbank Performance nach OraBench T132 PL/SQL Arithmetic Mix Datatype Number

	IBM Power7 2 sockets 3.5 GHz 12 cores, 48 threads Oracle Lizenz 12			Intel E5540 4 sockets, 2.53 GHz 16 cores, 32 threads Oracle Lizenz 8		
	CPU [%]	Durchsatz Total [kOps]	Durchsatz Prozess [kOps]	CPU [%]	Durchsatz Total [kOps]	Durchsatz Prozess [kOps]
1	2.8	339	339	5.1	512	512
2	4.7	678	339	6.2	1'000	500
4	8.3	1'356	339	6.1	2'000	500
8	15.0	2'581	323	22.3	3'809	476
16	25.6	3'137	196	48.1	8'421	526
32	55.4	4'211	132	92.6	10'491	327
64	87.2	4'812	75	96.2	10'847	169

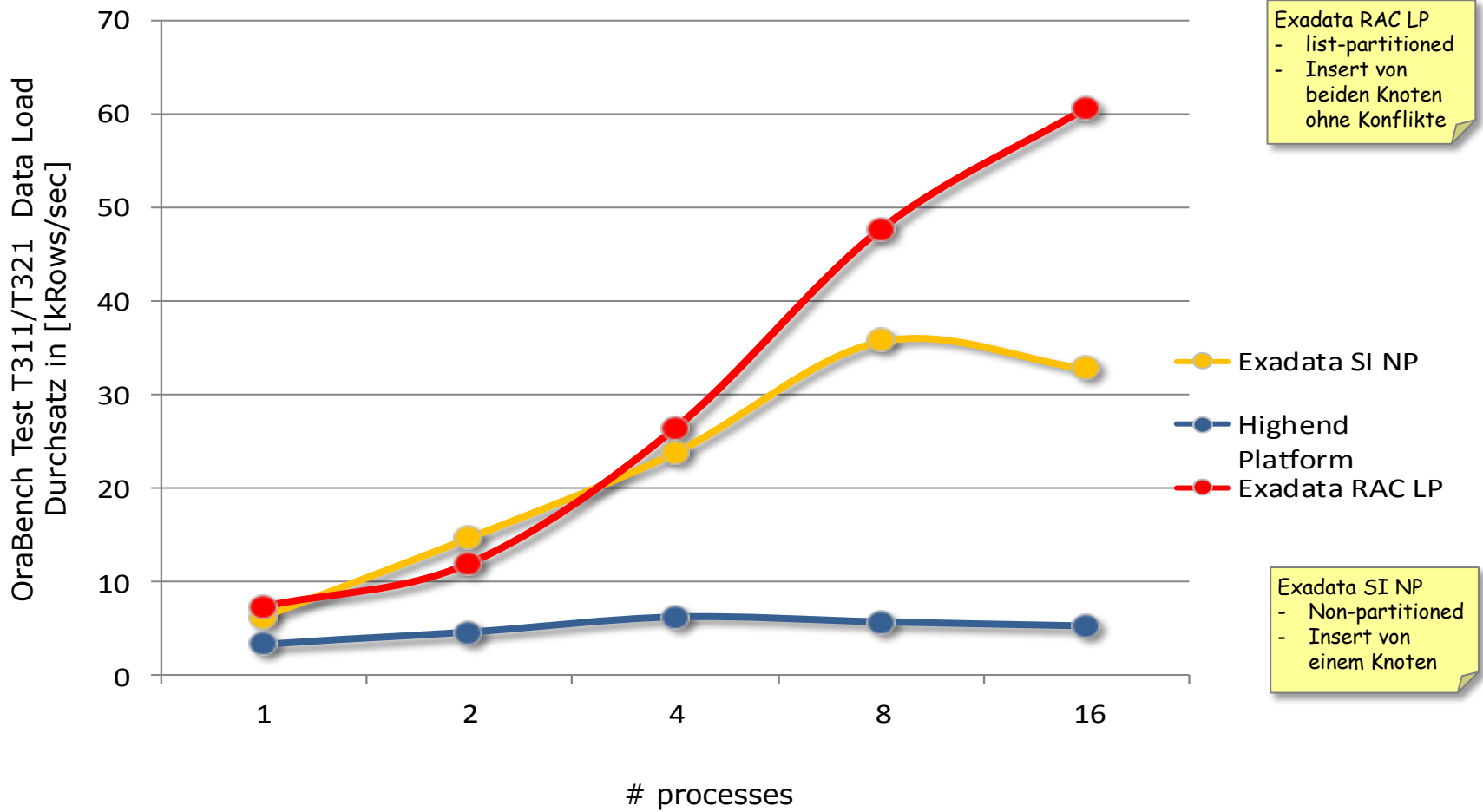
# Exadata Storage Server Storage Performance Sequential Write



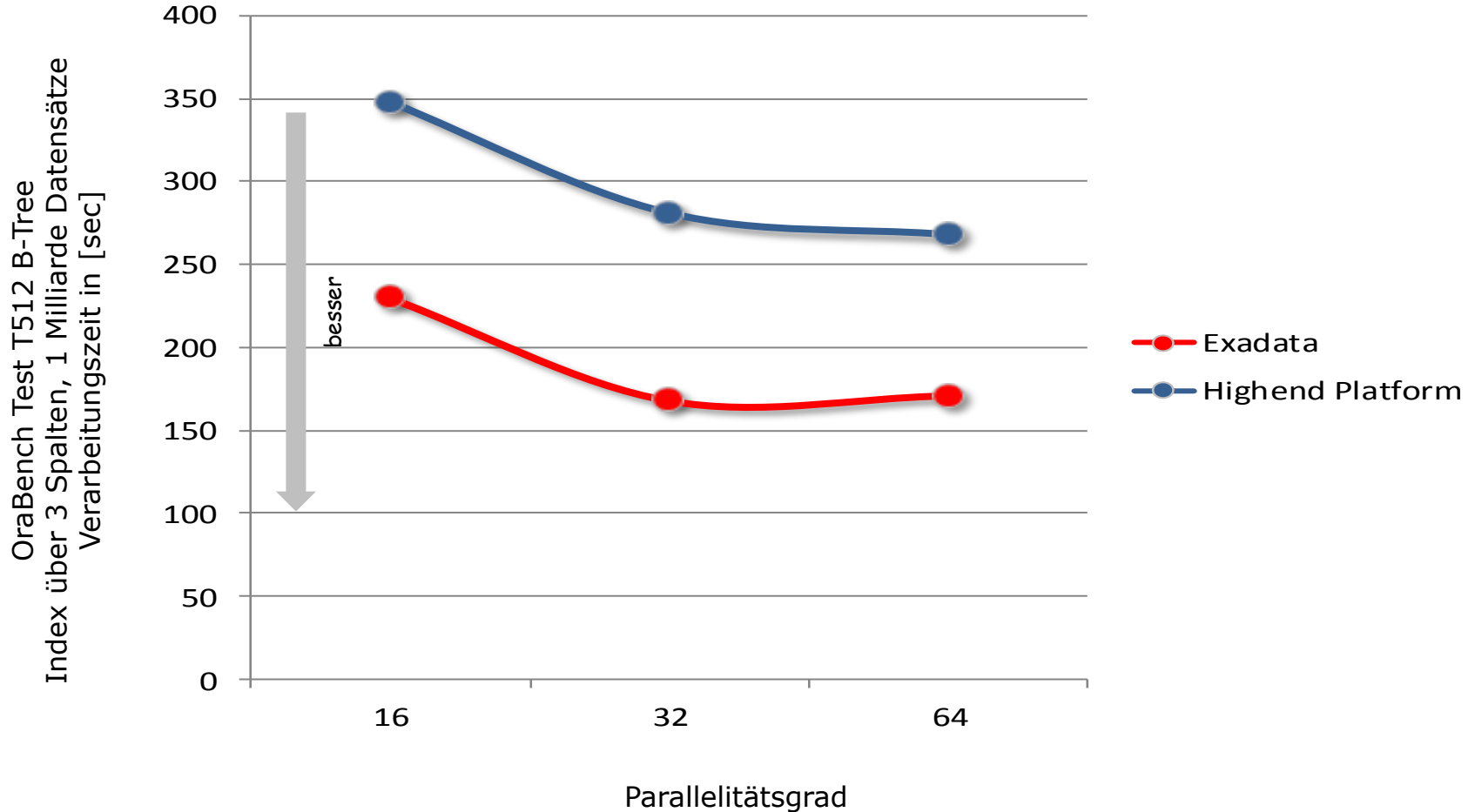
# Exadata Storage Server Storage Performance Sequential Read



# Exadata Database & Storage Server Performance Transactional Load



# Exadata Database & Storage Server Performance Data Aggregation



# Flash Technologie



	Disk Technologie	Flash Technologie
Modell	Seagate Cheetah	Sun Flash F20
Kapazität	600 GByte	96 GByte
Listenpreis	~ 750 USD	~ 4'700 USD
Random I/O (read)	300 IOPS	20'000 IOPS
Servicezeit	< 10 msec	< 1 msec
Sequential I/O (read)	100 MBps	1'000 MBps
Kosten pro GByte	1.25 USD/GByte	50.00 USD/GByte
Kosten pro IOPS	2.50 USD/IOPS	0.25 USD/IOPS
Kosten pro MBps	7.50 USD/MBps	4.70 USD/MBps
Kosten pro IOPS pro GByte (Agilität)	1'500 USD per IOPS/GByte <sup>1)</sup>	23.5 USD per IOPS/GByte

<sup>1)</sup> Bei Verwendung von SATA Disks wäre diese Kennzahl deutlich höher (und damit schlechter)



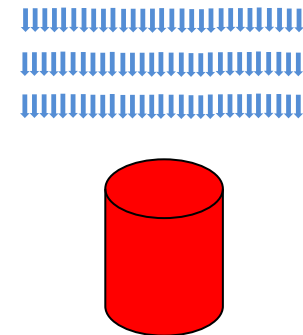
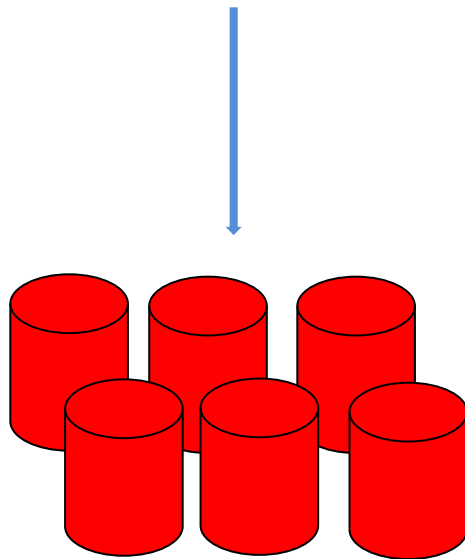
# Bewirtschaftung von Storage

## ■ HDD

- 600 GByte
- 300 IOPS
- < 10 msec

## ■ Flash

- 96 GByte
- 20'000 IOPS
- < 1 msec





- Verschiedene Cache Optionen mit Oracle 11.2

```
SQL> alter table test storage (cell_flash_cache keep); -- nur Exadata
```

Table altered.

```
SQL> alter table test storage (flash_cache keep);           -- nur 11.2 mit  
                                                            Solaris und OEL
```

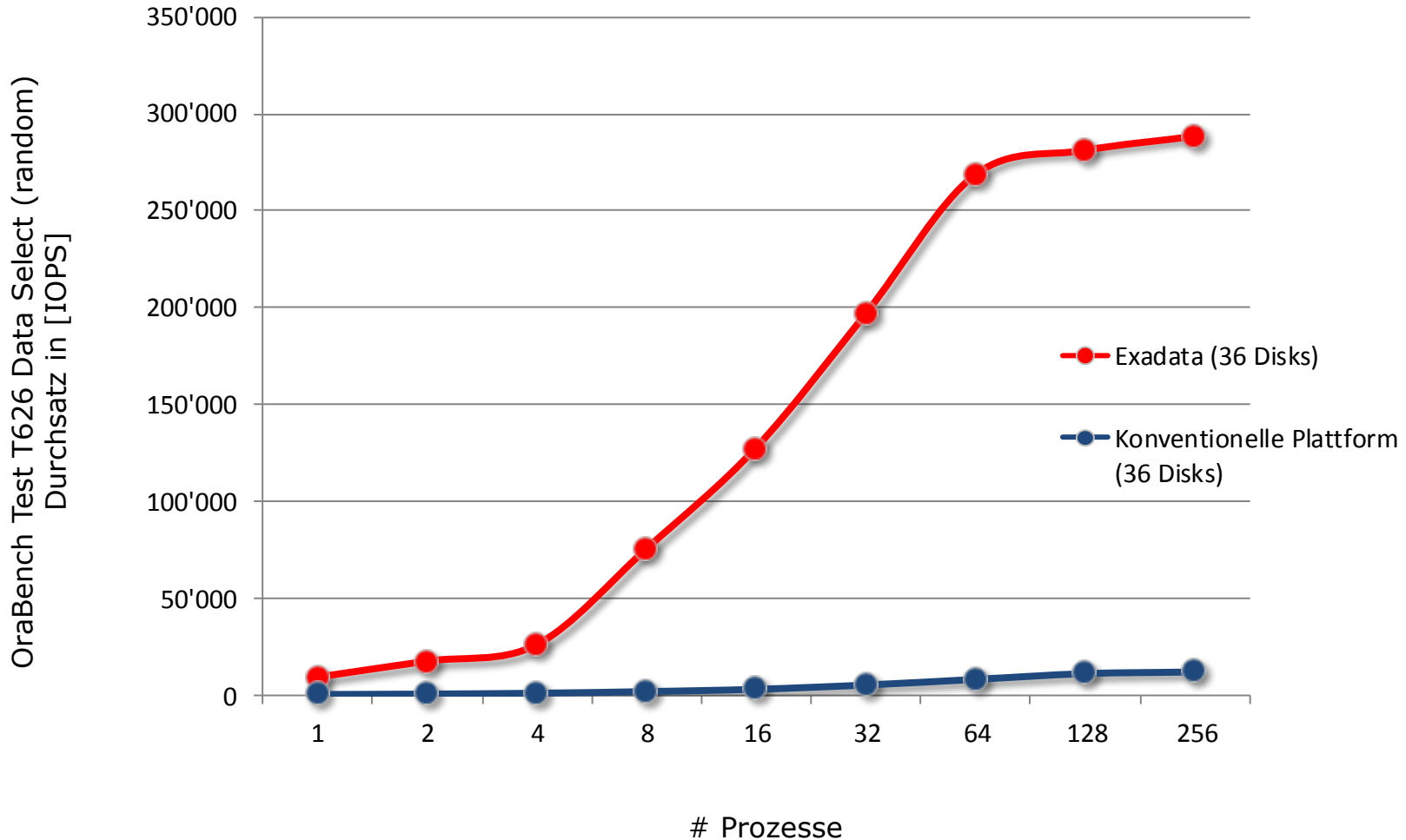
Table altered.

```
SQL> alter table test cache;
```

Table altered.

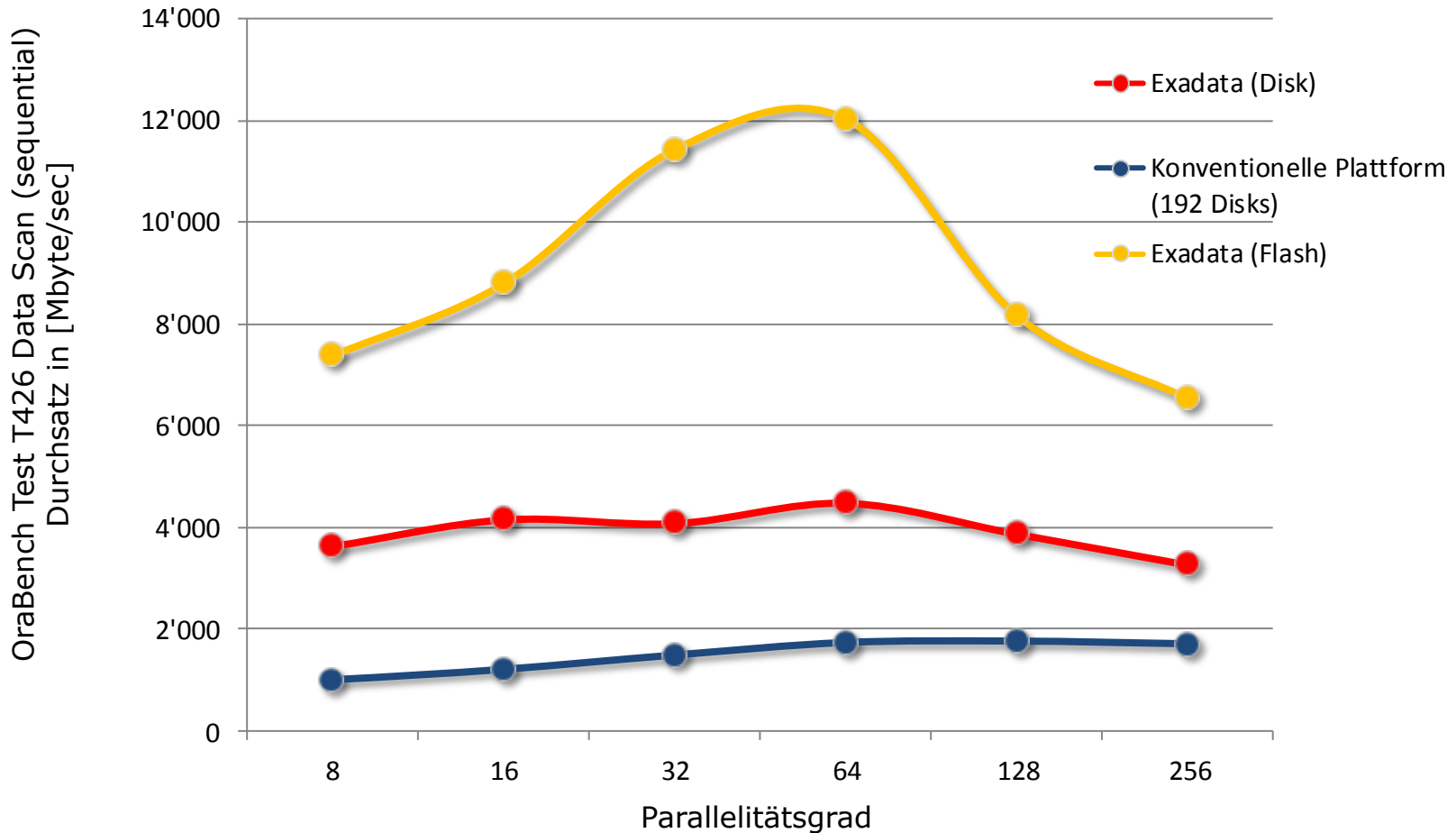
# Exadata Flash Technologie

## Performance Flash Cache Random I/O



# Exadata Flash Technologie

## Performance Flash Cache Sequential I/O



# Arbeiten mit der Exadata



- Dokumentation leider nicht allgemein verfügbar:
  - *Oracle Exadata Storage Server Users Guide 11g Release 2; E13861-04, December 2009*
  - *Oracle Exadata Storage Server Software Release Notes 11g Release 2, E13862-03, November 2009*
  
- Neue Konfigurationsparameter
- Neue Statistikwerte
- Neue Wait Events
- Neue Execution Plans

# Arbeiten mit der Exadata



- **Empfehlung**

- Hands-on Workshops mit erfahrenen Exadata Consultants in Exadata Testcenter
- Mehr Informationen z.B. im Schweizer Exadata Testcenter der Tradeware

**Tradeware**  
member of the LAKE GROUP  
[www.exadata.ch](http://www.exadata.ch)

# Proof-of-Concept?

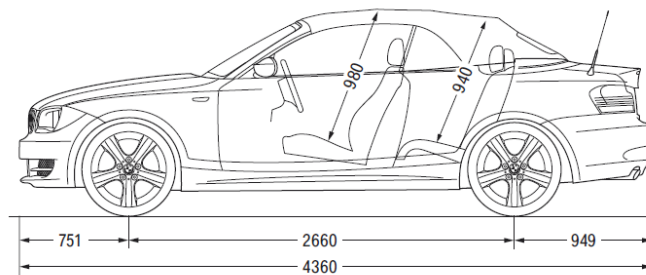
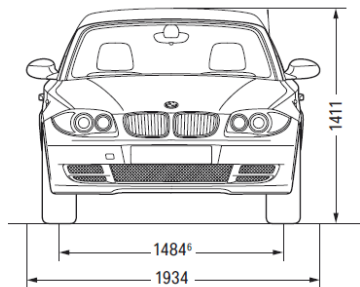


## Engine

Cylinders/valves	4/4
Capacity in ccm	1,995
Stroke/bore in mm	90.0/84.0
Max. output in kW (hp) at 1/min	105 (143)/6,000
Max. torque in Nm at 1/min	190/4,250
Power-to-weight ratio (EU) in kg/hp	10.5

## Performance

Drag (cw)	0.32
Top speed (km/h)	210
Acceleration 0 - 100 km/h (in s)	9.3
Acceleration 0 - 1,000 m (in s)	30.6
Acceleration 80 - 120 km/h in 4th/5th gear (in s)	9.6/12.5



# Proof-of-Concept?



- Äusserst aufwendig (mehrere Mannwochen)
  - Vorbereitung
  - Durchführung
  - Auswertung
  - Anonymisierung sensibler Daten
  
- Beispiele
  - Simulation von OLTP Anwendungen (ev. mit RAT)
  - Konsolidierung von Oracle Plattformen

# Proof-of-Concept?



- Benötigt skalierbare Applikationen
- Ergebnisse häufig nicht reproduzierbar
  - Bei Applikationsänderungen
  - Bei verändertem Datenbestand
- Wir empfehlen für den Vergleich der Exadata mit konventionellen Plattformen
  - verständliche
  - repräsentative
  - nachvollziehbarePerformance Kennzahlen für typische Datenbank Operationen

# Zusammenfassung (1)



- Database Server
  - x86 Prozessoren bieten bestes Preis-/Leistungsverhältnis
- Kennzahlen zum Preis-/Leistungsvergleich DB Server
  - CPU Leistung pro USD
  - RAM Kapazität pro USD
  - Leistung pro Oracle Lizenz USD

# Zusammenfassung (2)



- Storage Server
  - Extreme I/O Performance (I/O Tsunami)
  - Verlangt enge Kopplung von Storage- und Datenbank-Server
  - Kann mit konventionellen Architekturen – wenn überhaupt – nur mit grossem Aufwand nachgebaut werden
  - ASM als Volume Manager und File System liefert optimale Performance bei niedrigem Aufwand

# Zusammenfassung (3)



- Kennzahlen zum Preis-/Leistungsvergleich Storage Server
  - Kosten pro GByte (*capacity*)
  - Kosten pro IOPS (*random access pattern*)
  - Kosten pro MBps (*sequential access pattern*)
  - Kosten pro IOPS pro GByte (*storage agility*)
  
- Zu beachten
  - Storage Servicezeiten
  - Storage Area Network integriert
    - o Quarter Rack Konfiguration ausreichend für eine Leistung von 300'000 IOPS (*random I/O*) und 12.5 GByte/sec (*sequential I/O*)
  - Starke Kapazitätsreduktion durch Komprimierung
    - o Faktor 10 ist realistisch

# Zusammenfassung (4)



- Exadata liefert extreme Performance und hohe Verfügbarkeit out-of-the-box ... verglichen mit konventionellen Oracle Plattformen
  - Kein Engineering
  - Kein Storage Tuning
  - Kein Server Tuning
  - Kein Database Tuning
- Verantwortungsbereiche klären
  - Klassische Trennung von Storage-, Server- und Datenbank Betrieb funktioniert nicht mehr ...

# Zusammenfassung (5)



- Write-through Flash Cache Technologie
  - Hervorragend bei Abfragen
  - Automatischer Cache Algorithmus funktioniert schnell und wirkungsvoll
  - ARCH-, DBWR- und LGWR-Prozesse profitieren nicht vom Flash Cache
  - Kein Problem bei Data Insert, aber Data Update (DBWR Prozess)

# Zusammenfassung (6)



- Storage Index äusserst nützliches Feature
  - Nicht-indizierte Queries auf Tabellen (225 GByte) in weniger als 10 Sekunden
  - Keine *side effects*
- Exadata kann *mixed-workloads* verarbeiten
  - Ressource Management
  - Wichtig für Server Konsolidierung

# Zusammenfassung (7)



- HCC (*hybrid columnar compression*)
  - Konnte in unserem Test Daten von 240 GByte auf 17 GByte komprimieren (*query high mode*)
  - Aber extrem CPU lastig
  - Läuft nicht auf dem Storage Server, sondern auf dem Database Server
  - Wünschenswert: *compliance mode* für Archivierung

<sup>2)</sup> Mehr Details in DOAG 2010 Konferenzvortrag von Martin Bracher: Oracle Advanced Compression

IT Consulting & Engineering

# Effiziente und extrem schnelle Oracle Plattformen. Ihr Vorteil. Unsere Passion.

In&Out AG IT Consulting & Engineering  
Kilchbergsteig 13 CH-8038 Zürich  
Helvetiastrasse 5 CH-3005 Bern

Phone Zürich +41 44 485 60 60  
Phone Bern +41 31 352 32 32  
Fax +41 44 485 60 68

[info@inout.ch](mailto:info@inout.ch), [www.inout.ch](http://www.inout.ch)



**in&out**



## **Manfred Drozd**

Dipl.-Inform.

Manfred Drozd studierte Informatik an der Universität Paderborn (Deutschland). Er hat die Entwicklung der relationalen Datenbanktechnik von Anfang an miterlebt und begann seine Karriere 1980 als Programmierer einer relationalen Datenbank. Von 1984 bis 1986 war er bei einem Basler Pharmazieunternehmen für die Einführung von Oracle Version 3.1 in deren technisch-wissenschaftlichem Rechenzentrum verantwortlich. Während dieser Zeit unterrichtete er als nebenamtlicher Dozent die Fächer Computer Architektur und Datenbanksysteme an der HTL in Bern und Basel. Zwischen 1986 und 1990 leitete Manfred Drozd verschiedene Datenbankentwicklungsteams. Von 1990 bis 2001 war er für die Firma Oracle (Schweiz) tätig, zuletzt als Direktor der Consulting Practice *Server Technology & Performance Architecture*. Heute arbeitet Manfred Drozd als unabhängiger Consultant und unterstützt Oracle Anwender bei Design, Implementierung und Optimierung von Oracle Plattformen.

Seit 1995 beschäftigt sich Manfred Drozd schwerpunktmässig mit dem Thema Oracle Performance. Er führt regelmässig Performance Tests in den Benchmark Centers der Hardware Hersteller durch und leitet Seminare zum Thema Performance Tuning. Er ist häufiger Referent bei SOUG (Swiss Oracle User Group) und DOAG (Deutsche Oracle Anwendergruppe) Veranstaltungen. Manfred Drozd und sein Team entwickelten OraBench zur Ermittlung von Performance Kennzahlen von Oracle Plattformen. OraBench Kennzahlen helfen, dass Performanceverhalten von Plattformen auf der Basis von Fakten zu verstehen.

Manfred Drozd ist Anhänger des ganzheitlichen *Performance by Design* Ansatzes: Oracle Database Plattformen (Towers) werden mit einer optimalen Abstimmung aller Layer aufgebaut, um Performance- und Verfügbarkeitsanforderungen von Anwendungen effizient zu implementieren. Er setzt diese Methode äusserst erfolgreich bei der Architektur grosser OLTP und Data Warehouse Systeme in der Telekommunikations- und Finanzindustrie ein.