Daniel Hillinger

Database Administrator

VALUE TRANSFORMATION SERVICES
an IBM subsidiary

Hugepages, NUMA or nothing on Linux?

Value Transformation Services

VALUE TRANSFORMATION SERVICES
an IBM subsidiary

1. September 2013 gegründet

Joint Venture zwischen IBM und Unicredit

- 1000 Mitarbeiter
- 6 Ländern
- 6 Rechenzentren

www.linkedin.com/company/value-transformation-services

Agenda

- Memory Grundlagen
- NUMA
- Hugepages
- Kompatibilität
- Empfehlungen
- Performance Test
- Vor- und Nachteile
- Exadata

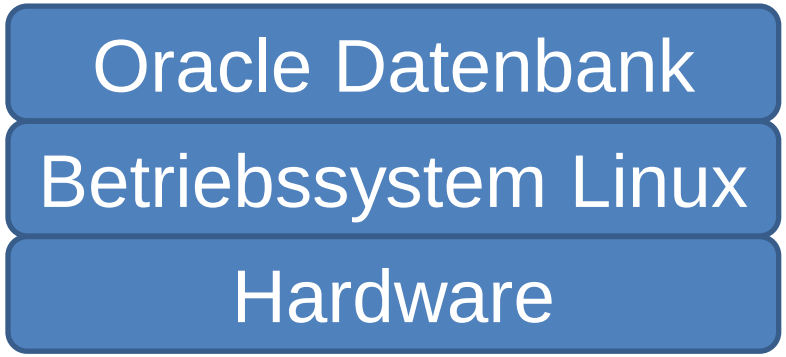
Memory Grundlagen

Flüchtiger Speicher

Sehr schneller Speicher

SGA wird direkt bei Start allokiert

PGA je nach Bedarf im laufenden Betrieb

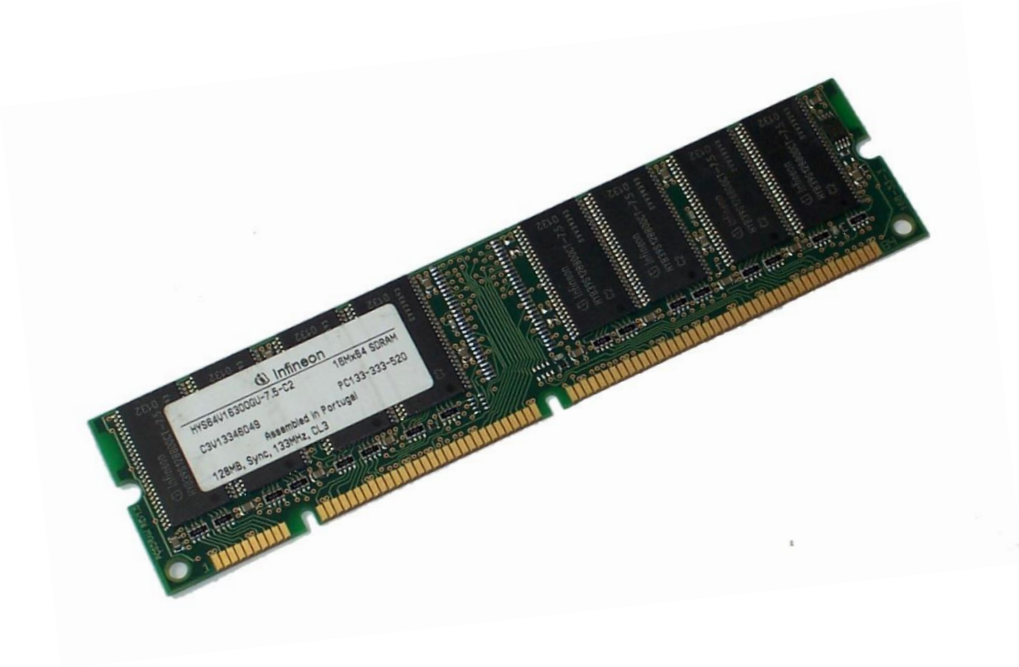


Wie wird Memory verwaltet

Standardmäßig wird Memory in 4KB großen Blöcken verwaltet

Ältere Systeme:
Memory Controller ist eine eigene Komponente

Aktuelle Systeme:
Memory Controller ist in die CPU integriert

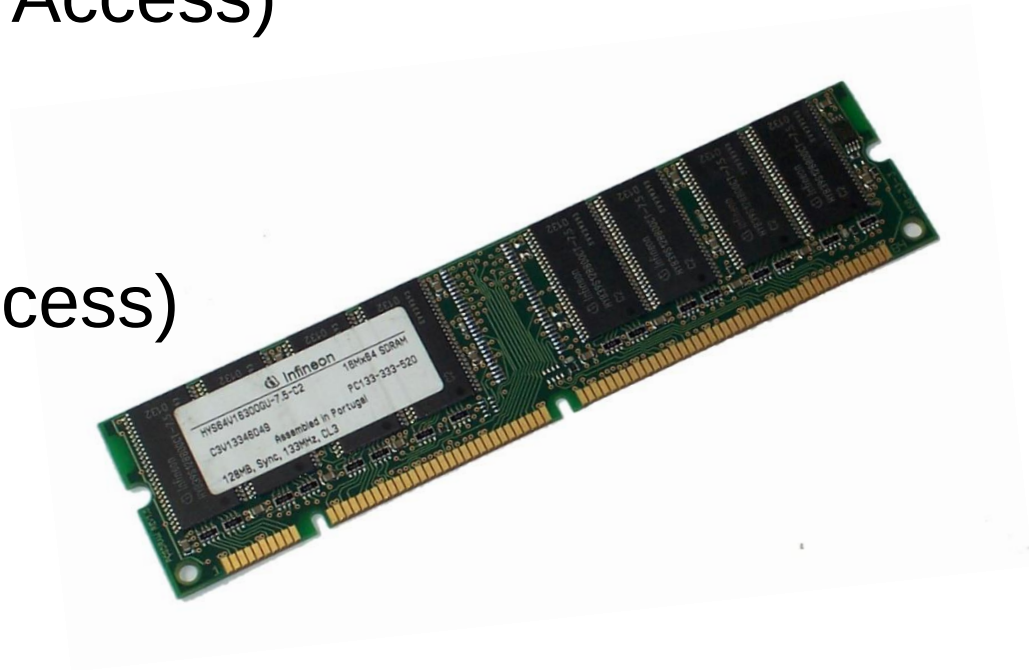


Wie wird Memory verwaltet

Wie wird der Arbeitsspeicher bei Multi-CPU Systemen angesprochen?

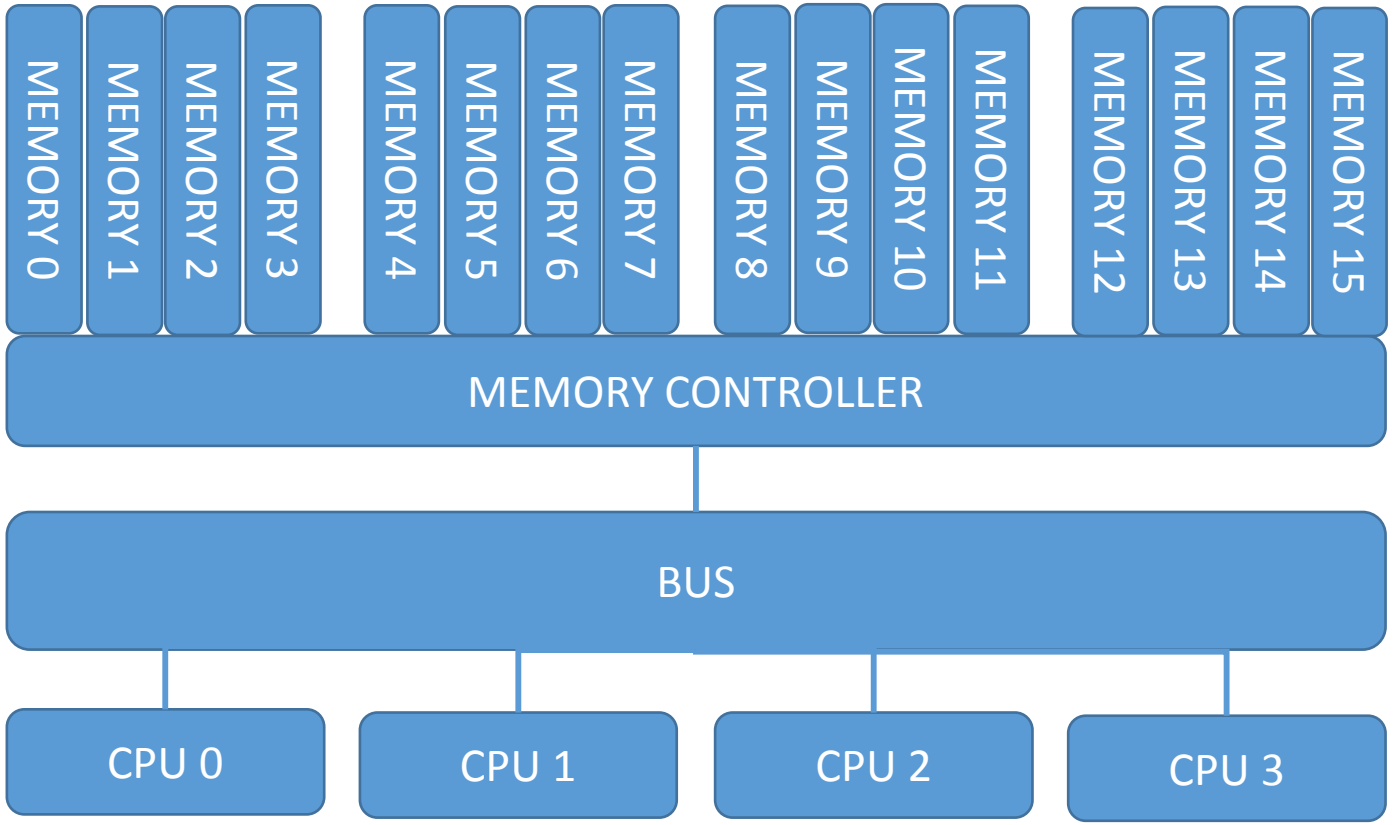
- Symmetric Multi-Processor (SMP) Architektur
  →UMA (Uniform Memory Access)

- NUMA
  (Non Uniform Memory Access)



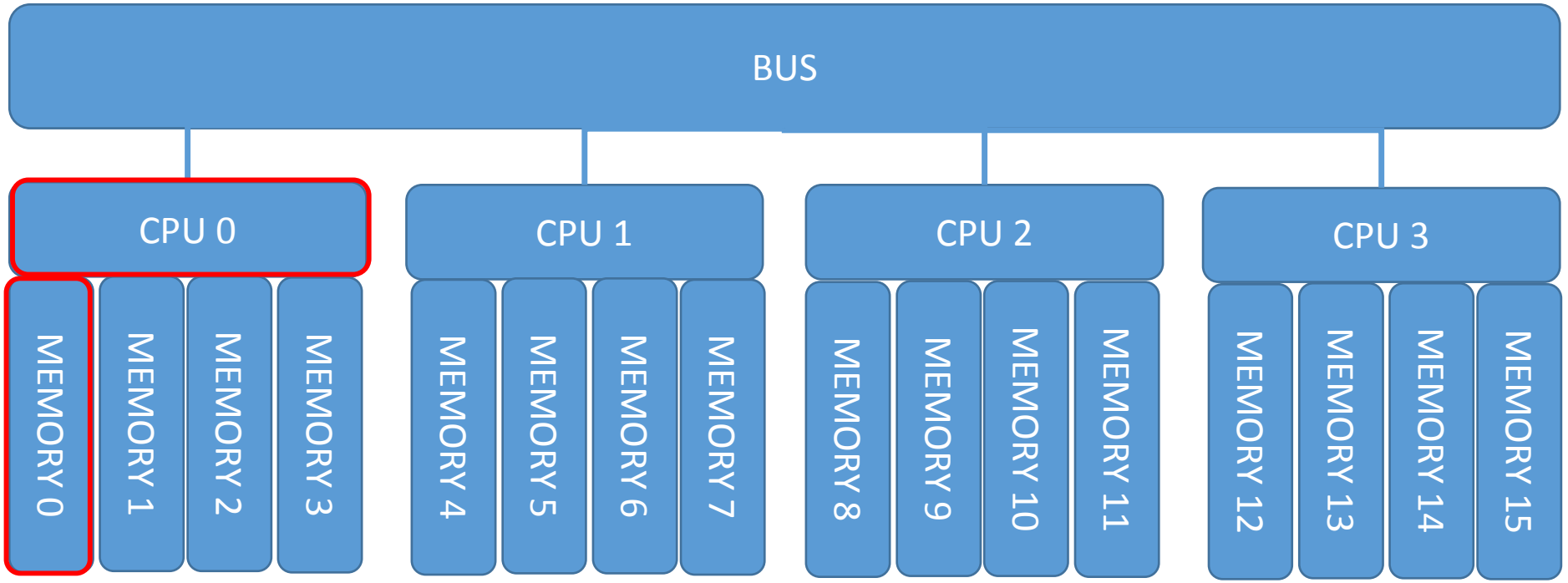
UMA

Jede CPU hat die gleichen Zugriffszeiten zu jedem Memory DIMM



NUMA

CPUs haben unterschiedliche Zugriffszeiten je nachdem ob der Memory local oder remote ist.



NUMA

NUMA ist eine physikalische Memory Architektur
Nur bei Multi-CPU Systemen möglich

Hardware, Betriebssystem und Applikation müssen NUMA unterstützen

Es wird die CPU verwendet, die dem benötigten Memory am nächsten ist

Ist mein System NUMA-fähig?

```
$ numactl --hardware
available: 4 nodes (0-3)
node 0 cpus: 0 1 2 3 4 5 6 7 8 9 40 41 42 43 44 45 46 47 48 49
node 0 size: 262122 MB
node 0 free: 175478 MB
node 1 cpus: 10 11 12 13 14 15 16 17 18 19 50 51 52 53 54 55 56 57 58 59
node 1 size: 262144 MB
node 1 free: 176411 MB
node 2 cpus: 20 21 22 23 24 25 26 27 28 29 60 61 62 63 64 65 66 67 68 69
node 2 size: 262144 MB
node 2 free: 175004 MB
node 3 cpus: 30 31 32 33 34 35 36 37 38 39 70 71 72 73 74 75 76 77 78 79
node 3 size: 262144 MB
node 3 free: 174307 MB
node distances:
node   0   1   2   3
  0:  10  11  11  11
  1:  11  10  11  11
  2:  11  11  10  11
  3:  11  11  11  10
```

NUMA

NUMA ist ab Linux Kernel 2.6.14 verfügbar, muss aber als Kernel Build Option angegeben worden sein

```
$ uname -r
2.6.32-279.el6.x86_64

$ numactl --show
policy: default
preferred node: current
physcpubind: 0 1 2 3 4 5 6 7 8 9 10 11
cpubind: 0
nodebind: 0
membind: 0
```

NUMA

Oracle Datenbank ab 10.2.0.4 und 11.1.0.7
ACHTUNG:

NUMA ist standardmäßig aktiviert
Patch 8199533 schaltet es wieder aus

SPFILE Parameter:

Parameter type	String
Syntax	_enable_NUMA_optimization = { TRUE | FALSE }
Default value	FALSE
Modifiable	No
Basic	No
Oracle RAC	Multiple instances can use different values

Hugepages

Standardmäßig 2MB groß

bis zu 256MB unterstützt

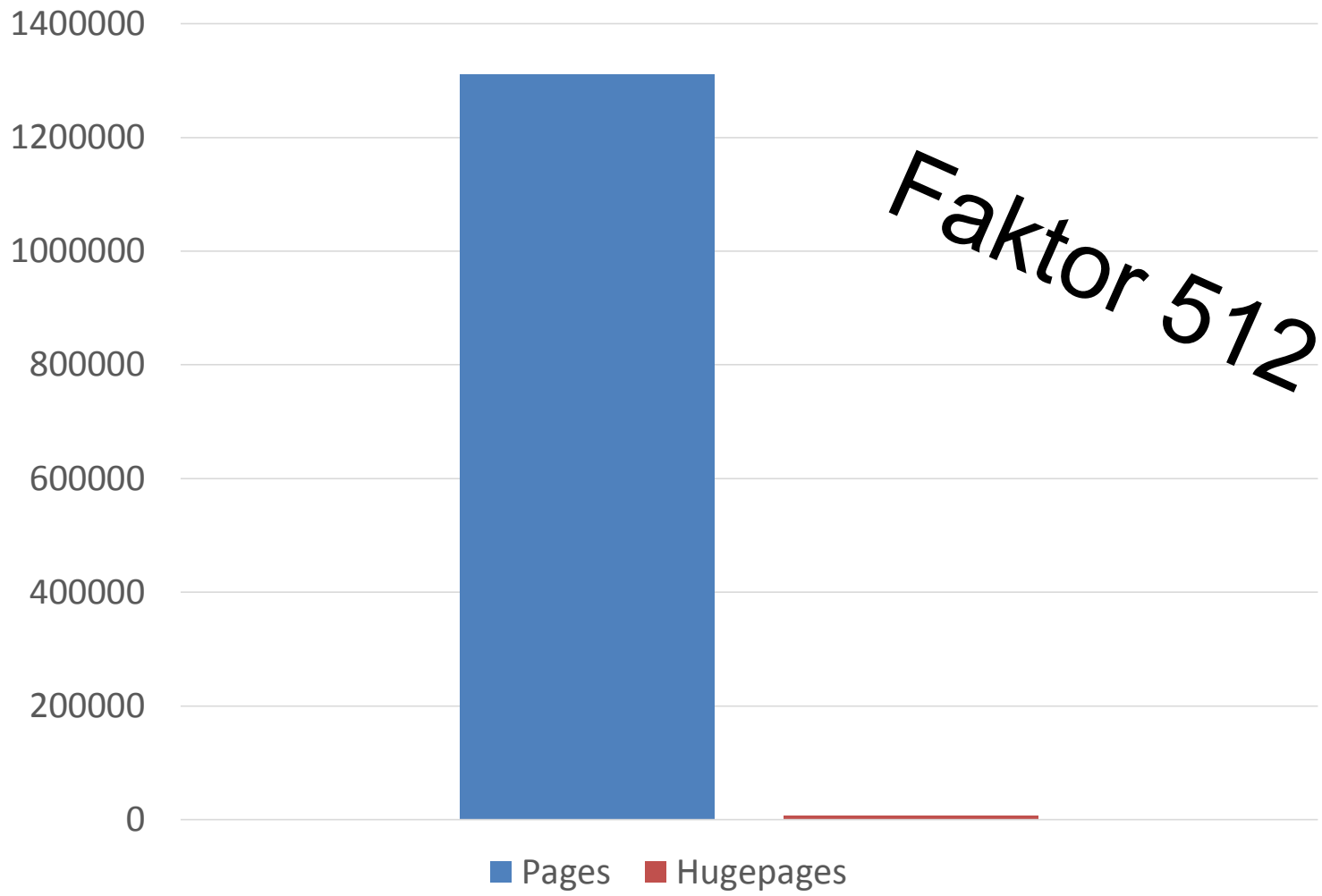
- Können nicht ausgelagert werden
- nicht vom Betriebssystem nutzbar

Vorteile:

- kswapd deutlich weniger beschäftigt
- Pagetable schneller durchsuchbar
- weniger Verwaltungsinformationen

Hugepages

Bei einer SGA von 5GB:



Hugepages

Verfügbar seit Kernel 2.6

Kernel Parameter:
nr_hugepages
hugepagesz

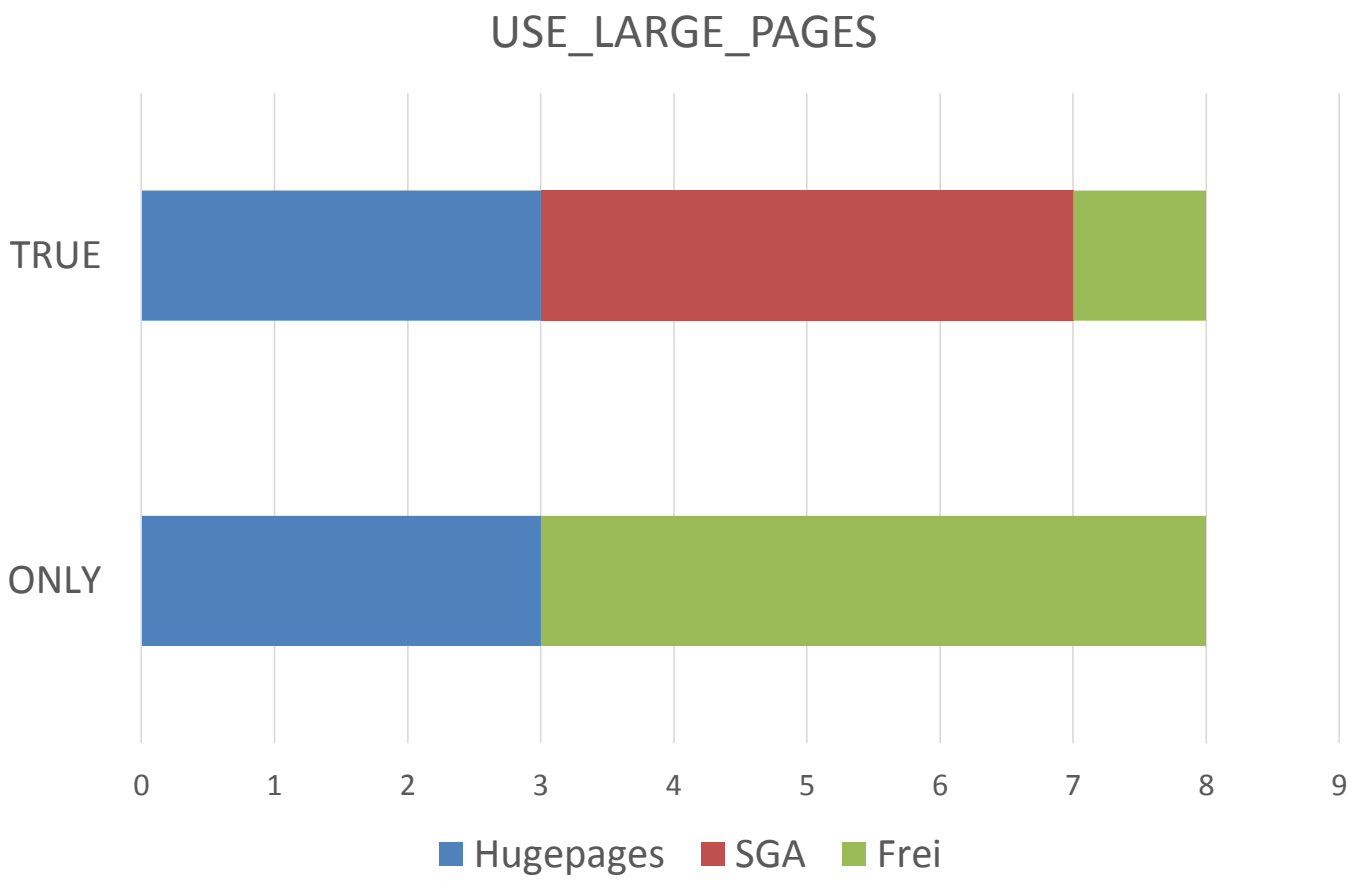
```
$ grep Huge /proc/meminfo
HugePages_Total:   150000
HugePages_Free:    149495
HugePages_Rsvd:       264
HugePages_Surp:         0
Hugepagesize:        2048 kB
```

Hugepages ab Oracle Version 11.2.0.0

SPFILE Parameter:

Parameter type	String
Syntax	USE_LARGE_PAGES = { TRUE | FALSE | ONLY }
Default value	TRUE
Modifiable	No
Basic	No
Oracle RAC	Multiple instances can use different values

Gefahr bei Hugepages



Hugepages

Version <= 11.2.0.2

```
****************** Huge Pages Information *****************
Huge Pages memory pool detected (total: 33280 free: 32222)
DFLT Huge Pages allocation successful (allocated: 20481)
***********************************************************
```

Hugepages

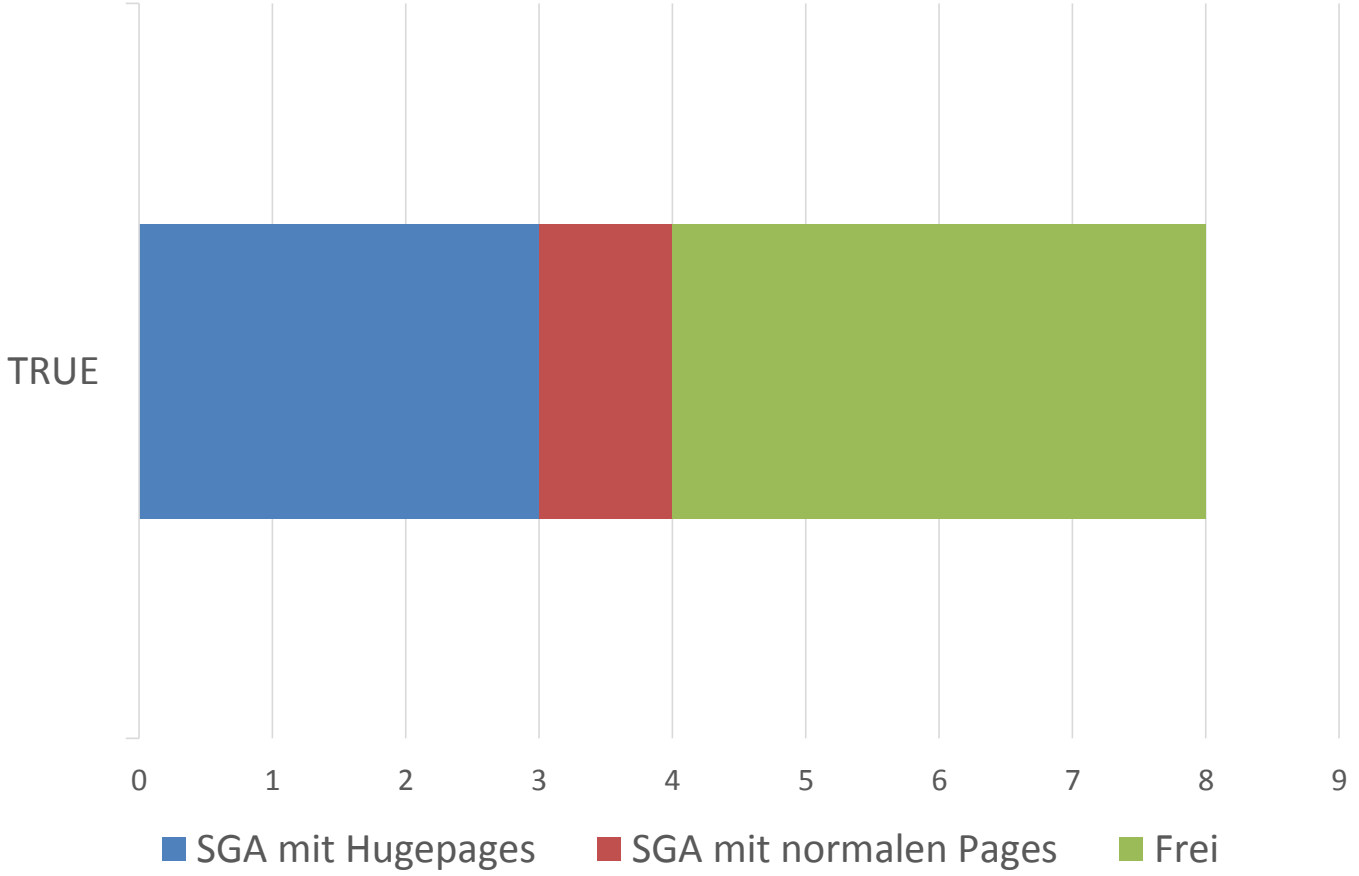
Version > 11.2.0.2

```
****************** Large Pages Information *****************

Total Shared Global Region in Large Pages = 20 GB (65%)

Large Pages used by this instance: 9985 (20 GB)
Large Pages unused system wide = 15 (30 MB) (alloc incr 64 MB)
Large Pages configured system wide = 10000 (20 GB)
Large Page size = 2048 KB

RECOMMENDATION:
  Total Shared Global Region size is 30 GB. For optimal performance,
  prior to the next instance restart increase the number
  of unused Large Pages by atleast 5361 2048 KB Large Pages (10 GB)
  system wide to get 100% of the Shared
  Global Region allocated with Large pages
***********************************************************
```

Hugepages

VALUE TRANSFORMATION SERVICES
an IBM subsidiary

Version > 11.2.0.2



Hugepages

Prüfen, ob man Hugepages online erweitern kann

(Spalte10 + 2 * Spalte11) * 2MB

```
$ cat /proc/buddyinfo
Node 0, zone      DMA      0      …      1      0      1      1      3
Node 0, zone    DMA32      9      …      9      9      9      5     83
Node 0, zone   Normal   6777      …    138     94    107      6  59896
Node 1, zone   Normal   1055      …    206    118    163     10  60343
Node 2, zone   Normal   7308      …    107    336    334      7  59817
Node 3, zone   Normal   2604      …    107    436    780     12  59811
```

Beispiel:

((6+10+7+12)+2*(59896+60343+59817+59811))* 2MB

= 959538MB = 937GB

Kompatibilität

Bei Oracle ist NUMA nicht mit Hugepages verträglich

PGA ist nicht mit Hugepages kompatibel

	NUMA	Hugepages
AMM	Ja	Nein
SGA + PGA	Ja	Ja (nur SGA)
manuell	Ja	Ja

Empfehlungen

Oracle empfiehlt:

NUMA muss mit großer Vorsicht eingesetzt werden

Sollte in einer Testumgebung ausführlich getestet sein

Meine Empfehlung:

Der Test sollte auf baugleicher Hardware stattfinden.
Beeinflussende Größen:

- Anzahl an NUMA Nodes
- Geschwindigkeit zwischen den CPUs

Empfehlungen

Oracle empfiehlt:

Hugepages sollten ab einer SGA Größe von 8GB verwendet werden

Meine Empfehlung ergibt sich aus dem Test …

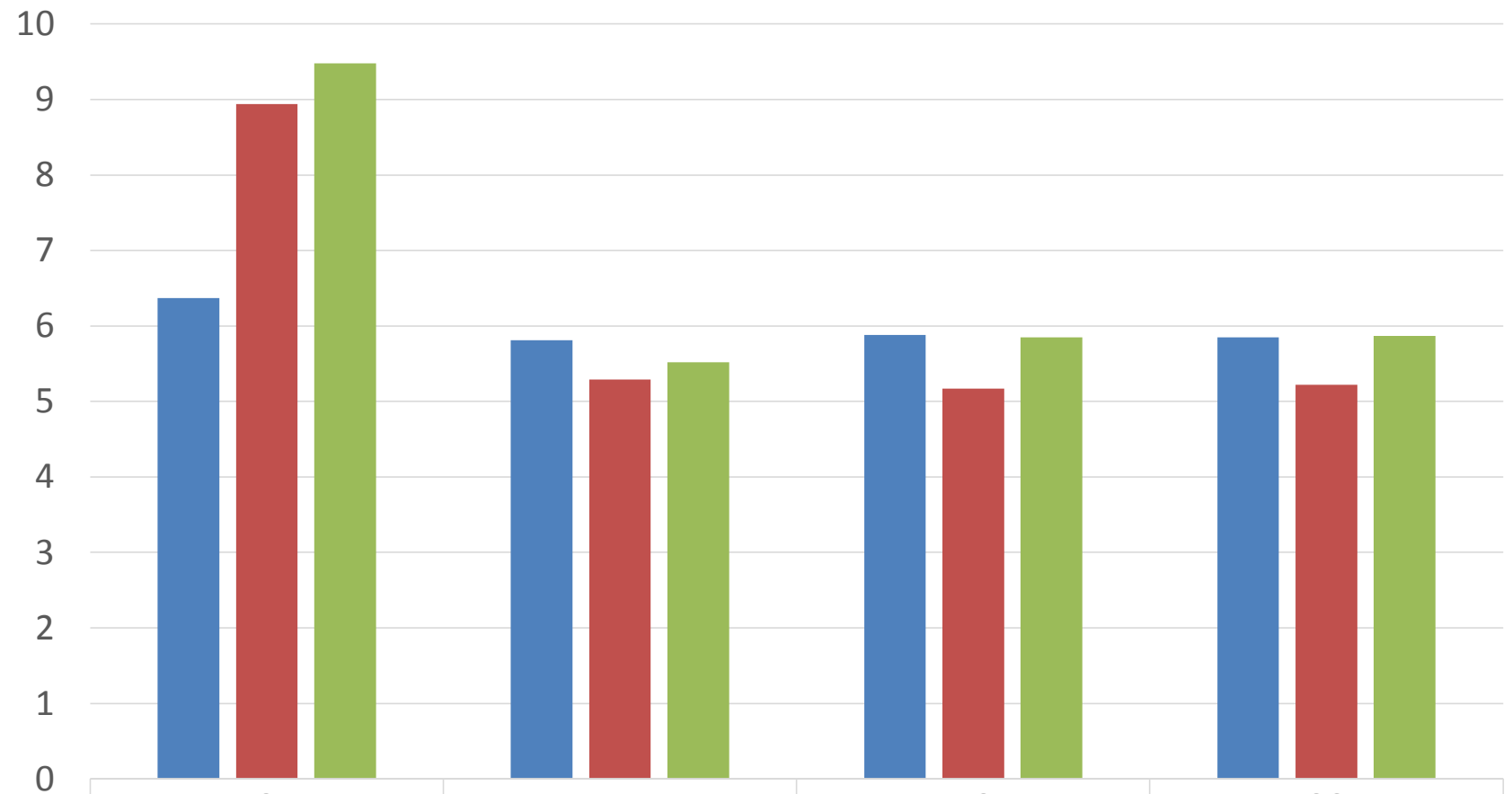
Performance

- 2GB große Tabelle
- Variierende SGA Größe 3G, 5G, 10G, 30G
- Daten sollen nur aus dem Memory gelesen werden

Testablauf:

- Neustart der Datenbank
- 3 Durchläufe
  select count(*) from test_tabelle;

Performance



	3	5	10	30
NUMA	6,37	5,81	5,88	5,85
Hugepages	8,94	5,29	5,17	5,22
nothing	9,48	5,52	5,85	5,87

Vor- und Nachteile NUMA

Vorteile

- (meistens) Performance Gewinn
- Kaum Konfiguration nötig
- Keine Einschränkung mit anderen Parametern

Nachteile

- Muss von Hardware unterstützt werden
- Oracle rät zur Vorsicht
- Unvorhersehbarere Performance

Vor- und Nachteile Hugepages

Vorteile

- Größerer Performance Gewinn nicht nur auf der Datenbank sondern auch für das Betriebssystem
- Stabilere Performance
- SGA kann nicht ausgelagert werden

Nachteile

- Nicht kompatibel mit AMM
- Zu kleinem Teil dynamisch erweiterbar
- Betriebssystem-parameter müssen angepasst werden

Exadata

Oracle‘s Antwort auf Performance Probleme?

Exadata

Die Exadata Hardware unterstützt NUMA nicht

Die Default Datenbank DBM verwendet Hugepages mit getrennt konfigurierter SGA und PGA

Fragen und Antworten…

ASM

Standardmäßig verwendet eine ASM Instanz AMM

Hugepages sind möglich, werden aber wegen dem geringen Speicherverbrauch nicht empfohlen

hugepages_settings.sh

- Oracle liefert ein Skript, mit dem man die optimale Anzahl an Hugepages ermitteln kann
- Vorsicht: Nur laufende Datenbanken werden berücksichtigt

```
$ ./hugepages_settings.sh
.
.
.
Recommended setting: vm.nr_hugepages = 22960
```