ORA-SOLUTIONS.NET

Mastering Oracle Performance, High Availability, Manageability

ORACLE 10g Certified Master

Unix Workshop for DBAs – Part 1: Storage

Munich, Oktober 2008

UNIX Workshop für DBAs – 3 Parts

1 Storage

- SAN/NAS
- RAID/SAME/ASM
- MSA/EVA
- Performance, Monitoring

2 Linux

- Booting, Netzwerk (Konfiguration, TCP/IP, Tracing, Bonding), Prozesse (Tracing: strace), VLM, I/O Scheduling, Packages, LVM, Raw Devices, Memory Management, Monitoring, cron, Kernel-Modules, SSH

3 HP-UX

- Memory Management, Kernel Parameters, Mount Options, LVM, Filesystem, Monitoring, Shell Scripting, Networking, APA, Oracle Specifics

Content – Part 1: Storage

A SAN (Storage Area Network) vs. NAS (Network Attached Storage)

B RAID / SAME (Stripe and Mirror Everything)

C ASM (Automatic Storage Management)

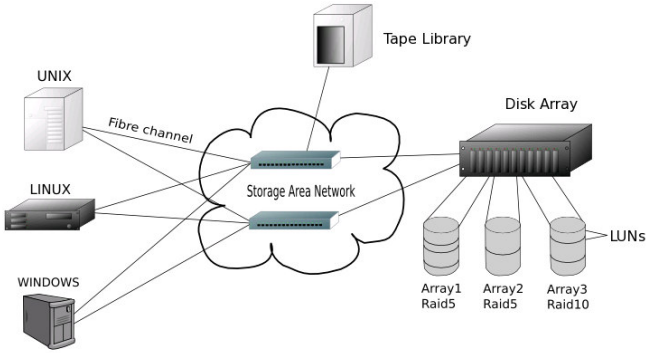
D MSA (Modular Storage Array)

E EVA (Enterprise Virtual Array)

F LVM (Logical Volume Manager)

G Performance-Monitoring / Durchsatz / Performance

SAN (Storage Area Network)



- block-basiert
- meist mit Fibre Channel realisiert
- Komponenten: FC-Switch, FC-Host-Bus-Adapter (HBA), SAN Storage
- FC-Adapter: 1 GBit/s, 2 GBit/s, 4 GBit/s
- Multi-Pathing
- bessere Performance als NAS
- LUNs (Logical Units) werden zur Verfügung gestellt (z.B. /dev/dsk/c4t0d2)
- Zoning definiert, welcher Fibre-Channel Switch Port, welche LUNs sehen darf

NAS (Network Attached Storage)

- file-basiert
- meist mit Gigabit Ethernet realisiert
- Protokoll: TCP/IP über GbE (NFS, SMB/CIFS)
- Multi-Pathing
- Mount-Points werden zur Verfügung gestellt (z.B. /mountpoint/filesystem_name)
- viele Storages sind sowohl NAS als auch SAN fähig
- z.B. NetApp Filer

RAID – Redundand Array of Inexpensive Disks

RAID0 (Striping):
Vorteil: Hohe Performance
Nachteil: keine Redundanz, wenn eine Platte im Stripe verbund ausfällt, sind die gesamten Daten aller Platten verloren.

RAID1: Mirroring
Vorteil: Read-Performance, Ausfallsicherheit.
Nachteil: Hohe Kosten, nur 50% der Plattenkapazität nutzbar.

RAID0+1: Striping + Mirroring
Zuerst wird über die Hälfte der Platten gestriped, dann wird das Stripe-Set auf die andere Hälfte der Platten gespiegelt.
Vorteil: Höhere Performance als bei purem RAID1
Nachteil: Fällt eine Platte aus, ist ein kompletter Mirror kaputt

RAID1+0: Mirroring + Striping (Vorteil gegenüber RAID0+1 bei Plattenausfall)
Vorteil: Performance-Vorteil gleich wie RAID0+1, bei Platten-Ausfall müssen nur die Daten von dieser Platte resynchronisiert werden.
Nachteil: Hohe Kosten pro MB im Vgl. zu RAID5. Nur 50% der Kapazität nutzbar.

RAID5: Striping mit Parity (http://www.baarf.com/, http://www.miracleas.com/BAARF/1.Millsap2000.01.03-RAID5.pdf)
Z.B. bei EVA: Ein Stripe Set beinhaltet 5 Disks, davon 1 als Parity. D.h., nur 80% der Kapazität sind nutzbar.
Vorteil: Geringere Kosten pro MB, gute Lese-Performance
Nachteil: Hoher Performance-Nachteil bei random writes (DBWR), Wenn kein ausreichend großer Storage Cache vorhanden ist, sind auch sequential writes (LGWR) problematisch.

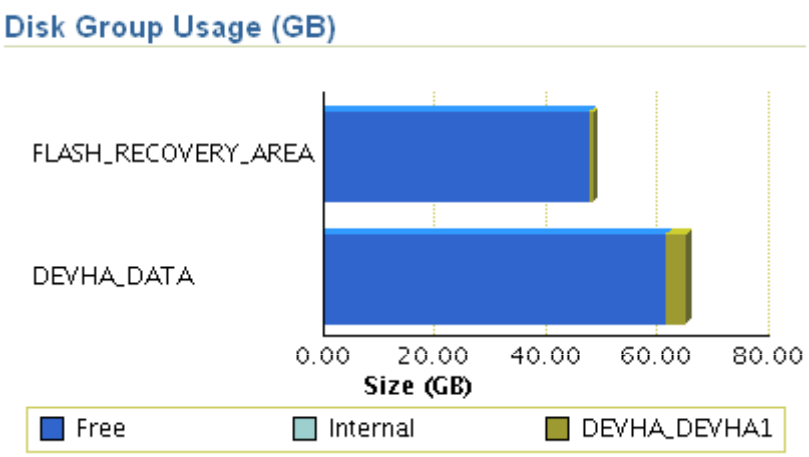
SAME (Stripe and Mirror Everything)

- Konzept von Oracle für maximale Datenbank-Performance
- Stripe across all available spindles
- then mirror all filesystems striped across all disks
- bei ASM realisiert

ASM (Automatic Storage Amanagement)

- Extent based load balancing mirroring
- Mittels ASMLib auch ohne Raw devices einsetzbar (LUNs oder Partitions)
- Ersetzt Logical Volume Manager und Filesystem
- Optimale Performance durch Async I/O und Direct I/O
- Begriffe:
  - DISK GROUP: beliebige Anzahl von Disks (oder LUNs oder Raw Devices) werden zu einer Disk Group zusammengefasst.
  - FAILURE GROUP: Eine Disk Group kann 1-3 Failure Groups beinhalten. Es müssen eigene Controller, eigene Platten pro Failgroups verwendet werden, um Single Point of Failures zu vermeiden.
    - Redundancy Normal: 2 Failgroups,
    - Redundancy High: 3 Failgroups,
    - Redundancy External: 1 Failgroup
- Feel free 2 test @ muc-dba04/muc-dba05

ASM (Automatic Storage Amanagement) (2) – Enterprise Manager



Disk Groups

TIP The usable free space specifies the amount of space that can be safely used for data. A value above zero means that redundancy can be properly restored after a disk failure.

Create | Mount All | Dismount All

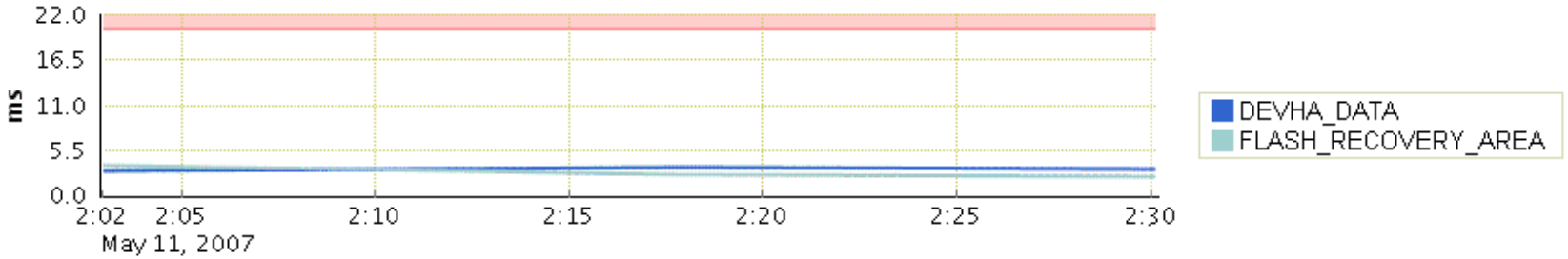
Delete | Actions | Mount | Go

Select	Name	State	Redundancy	Usable Free (GB)	Size (GB)	Used (GB)	Used (%)	Member Disks	Pending Operations
◉	DEVHA_DATA	MOUNTED	NORMAL	30.76	65.21	3.69	5.66	2	0
○	FLASH_RECOVERY_AREA	MOUNTED	EXTERN	47.79	48.43	0.64	1.33	1	0

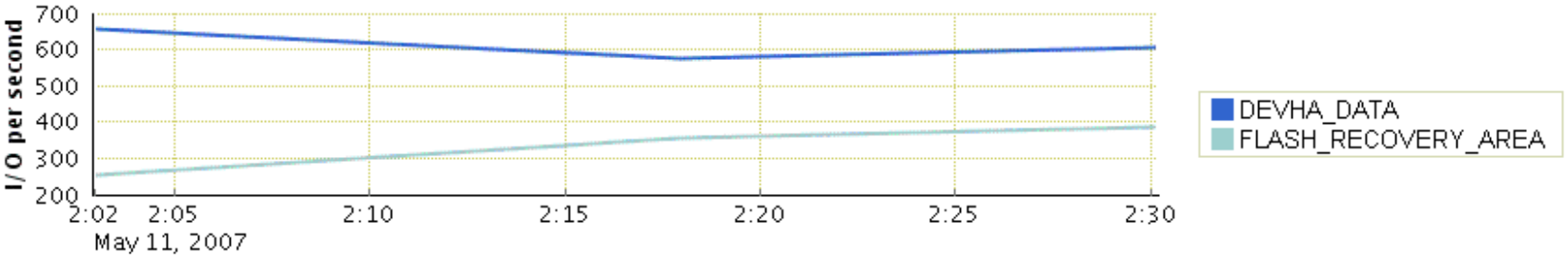
Home | Performance | **Administration** | Configuration

ASM (Automatic Storage Amanagement) (3) – Enterprise Manager

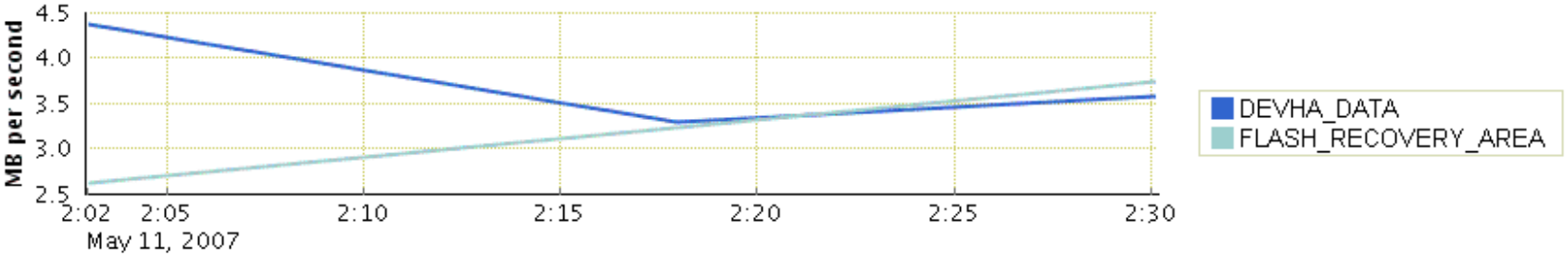
Disk Group I/O Response Time



Disk Group I/O Operations



Disk Group Throughput



MSA (Modular Storage Array)

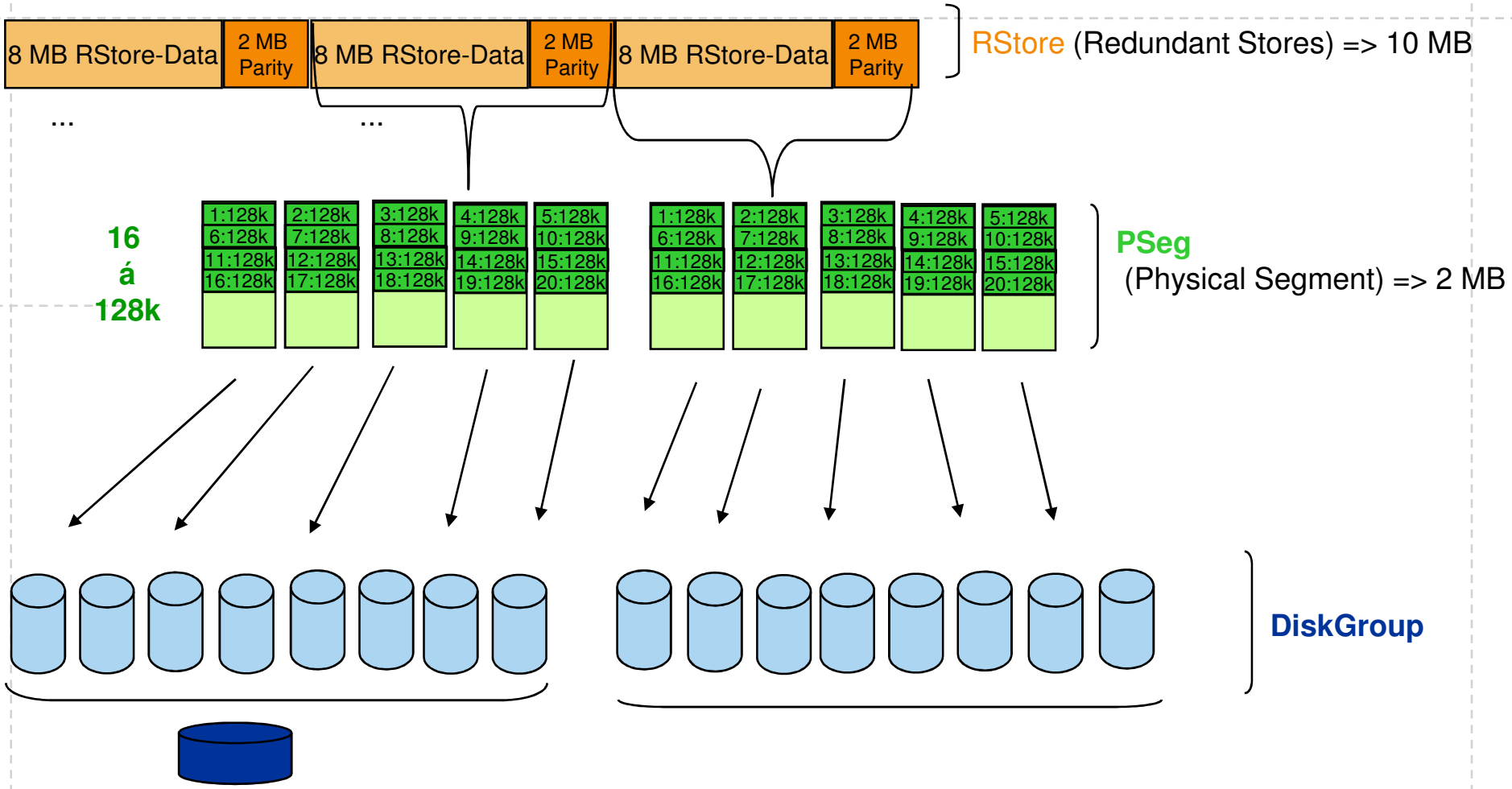
- kleinestes Storage System von HP
- Ausstattung:
  - Beschreibung: 145.6 GB Ultra320 SCSI Universal Festplatte (15.000 U/Min.)
  - Anzahl Platten: 14 je MSA
  - Anzahl MSA1000: 2
  - Anzahl MSA1000 Controller: 1 je MSA (Default, mehr sind von HP-UX nicht unterstützt)
  - Cache: max 512 MB, davon 50% für reads, 50% für writes
  - 64 kb Stripe Unit Size für RAID5
  - 5x72 GB plus 1 sparedisk

EVA (Enterprise Virtual Array)

- Mittleres Storage System von HP
- Facts:
  - 240 Disks max
  - Host Anschlüsse, 8
  - random reads: 54.000 I/Os pro Sec
  - sequential reads: 1430 MByte/sec
  - sequential writes: 525 MByte/sec
  - Supported Drives: 146GB 10K rpm,300GB 10K rpm, 72GB 15K rpm, 146GB 15K rpm, 300GB 15K rpm, 500GB FATA
  - Cache per Controller pair: max. 8 GB
  - RAID5 Implementation: 4+1 disks
- Begriffe:
  - Redundant Store (RStore), Physical Segment (PSeg), Redundant Storage Set (RSS)

EVA (Enterprise Virtual Array) (2)

LUN (=VDisk) /oradata



RSS (Redundant Storage Set)
6-12 disks => automat. Rebalancing

LVM (Logical Volume Manager) (1)

- Commands:
  - /usr/sbin/vgdisplay, /usr/sbin/pvdisplay, /usr/sbin/lvdisplay
- Freier Platz in Diskgroup (Achtung: RAID1: nur 50% Netto)

```
$ /usr/sbin/vgdisplay /dev/vg_mydb
--- Volume groups ---
VG Name                     /dev/vg_mydb
VG Write Access             read/write
VG Status                   available, shared, server
Max LV                      255
Cur LV                      80
Open LV                     79
Max PV                      128
Cur PV                      2
Act PV                      2
Max PE per PV               16383
VGDA                        4
PE Size (Mbytes)            32
Total PE                    6652
Alloc PE                    6614
Free PE                     38
```

LVM (Logical Volume Manager) (2)

- Problem mit Prefetch Mechanismus bei LVM Mirroring auf HP-UX
- Workaround: Schedule: sequential statt parallel
  - Liest nur von erstem Mirror
  - Schreibt hintereinander zuerst auf ersten Mirror, dann auf zweiten Mirror
  - Vorteil: Prefetching von Storage kann ausgenutzt werden
  - Nachteil: Write Zeit verdoppelt sich, wenn write Storage Cache zu klein

```
$ bdf
Filesystem           kbytes    used   avail %used Mounted on
/dev/vg_mydb/lvol6  2147450880 2122569016 24687480   99% /oracle/MYDB/oradata
/dev/vg_mydb/lvol7  1151336448 1127623376 23527816   98% /oracle/MYDB/oradata2

$ /usr/sbin/lvdisplay /dev/vg_mydb/lvol7
--- Logical volumes ---
LV Name                     /dev/vg_mydb/lvol7
VG Name                     /dev/vg_mydb
LV Permission               read/write
LV Status                   available/syncd
Mirror copies               1
Consistency Recovery        MWC
Schedule                    sequential   (default: parallel)
LV Size (Mbytes)            1124352
Current LE                  35136
Allocated PE                70272
```

LVM (Logical Volume Manager) (3)

- Problem mit Prefetch Mechanismus bei LVM Mirroring
  - oradata => **Lesen von 1. Storage Mirror**
  - oradata2 => **Lesen von 2. Storage Mirror**

```
/usr/sbin/lvdisplay -v /dev/vg_lmdwh/lvol6
--- Logical extents ---
  LE    PV1                 PE1   Status 1 PV2                 PE2   Status 2
  00000 /dev/dsk/c46t0d1    00900 current  /dev/dsk/c41t0d1    00900 current
  00001 /dev/dsk/c65t0d2    02948 current  /dev/dsk/c57t0d2    02948 current
  00002 /dev/dsk/c55t0d3    02948 current  /dev/dsk/c51t0d3    02948 current

/usr/sbin/lvdisplay -v /dev/vg_lmdwh/lvol7
--- Logical extents ---
  LE    PV1                 PE1   Status 1 PV2                 PE2   Status 2
  00000 /dev/dsk/c41t0d1    00901 current  /dev/dsk/c46t0d1    00901 current
  00001 /dev/dsk/c57t0d2    00900 current  /dev/dsk/c65t0d2    00900 current
  00002 /dev/dsk/c51t0d3    00900 current  /dev/dsk/c55t0d3    00900 current
```

Performance-Monitoring / Durchsatz / Performance (1)

- SAR

```
$ sar -d 1 100

HP-UX myhost B.11.23 U ia64    05/11/07

15:45:32   device   %busy   avque   r+w/s  blks/s  avwait  avserv
15:45:33   c2t0d0    4.90    0.50      10      86    0.00    9.24
           c2t1d0    6.86    0.50      14     102    0.00   10.09
           c4t0d2    0.98    0.50      12     120    0.00    1.79
           c7t0d2    1.96    0.50       6      71    0.00    4.59
15:45:34   c2t0d0    4.00    0.50       5      80    0.00   12.41
           c2t1d0    4.00    0.50       7      88    0.00   12.77
15:45:35   c2t0d0    2.02    0.50       4      65    0.00    7.46
           c2t1d0    4.04    0.50       6      73    0.00    7.76
           c4t0d2    2.02    0.50      14      83    0.00    1.74
           c7t0d2    2.02    0.50       7      26    0.00    3.41
15:45:36   c2t0d0    1.98    0.50       3      34    0.00    9.27
           c2t1d0    2.97    0.50       5      42    0.00    8.84
           c4t0d2    0.99    0.50      11     123    0.00    1.05
           c7t0d2    0.99    0.50       5      69    0.00    1.91
```

LUN	Auslastung	durchschn. Queue (>=0.5)	I/O/sec	512B/sec	Wait (Queue)	Service Zeit
						10-15 ms

Performance-Monitoring / Durchsatz / Performance (2)

- glance Advisor Script (60 Sek. Intervall)

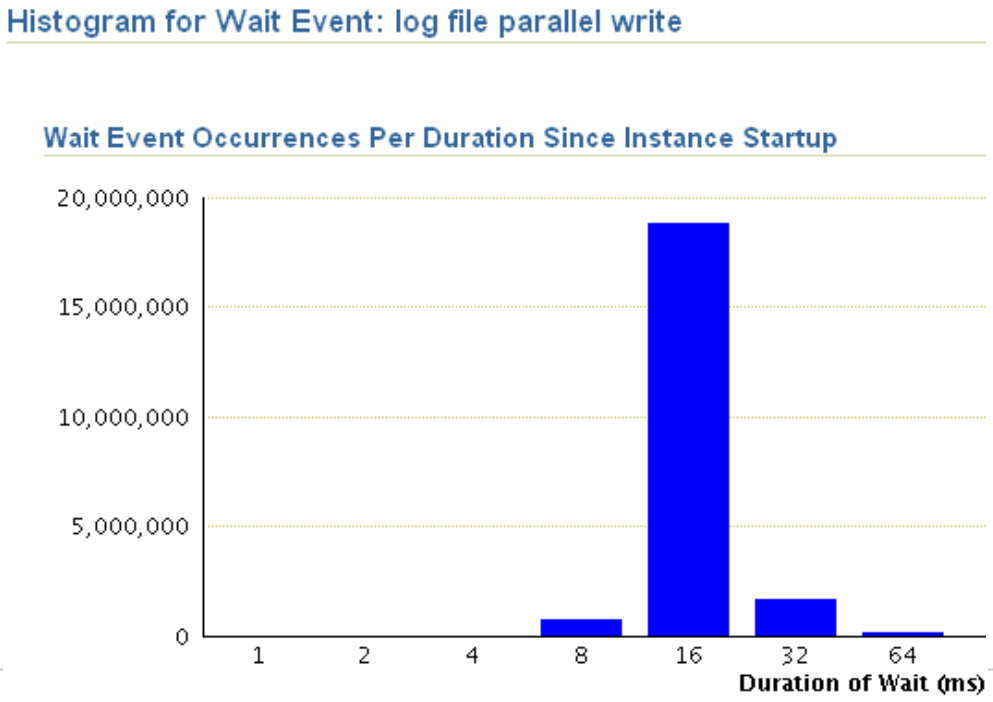
```
$ cat io2.conf
print gbl_stattime, gbl_cpu_total_util, gbl_disk_phys_read_rate ,
gbl_disk_phys_write_rate, gbl_disk_phys_read_byte_rate,
gbl_disk_phys_write_byte_rate

$ cat glance_script.sh
nohup glance -aos ./io2.conf -j 60 > glance_output_myhost_$$.txt
2>/dev/null &
```

```
10:14:33   52.5     4017.5        5.0   547794.1       30.0
10:15:32   36.5     3555.4       70.3   318020.5     2952.6
10:16:32   21.5     1021.3       94.2     8184.3     6689.4
10:17:32   35.2     1910.0       33.7   197938.1      955.3
10:18:32   58.5     1879.2      356.4   193373.5    27159.2
```

Zeit	CPU% (user+sys) -> ohne iowait	Read/s	Write/s	kB read/s	kB write /s

Performance-Monitoring / Durchsatz / Performance (2)

- Oracle Wait Events
  - db file sequential read (< 12 ms): Single-Block Disk Lesezugriff (Index Range / Unique Scan)
  - db file scattered read (< 15 ms): Multi-block Disk Lesezugriff (Table Scan, Index Fast Full Scan)
  - log file parallel write (0-2 ms) Wait-Event von LogWriter für Redo Log Write aus Log Buffer in Redo Log Files.



Fragen?

Martin Decker
www.ora-solutions.net