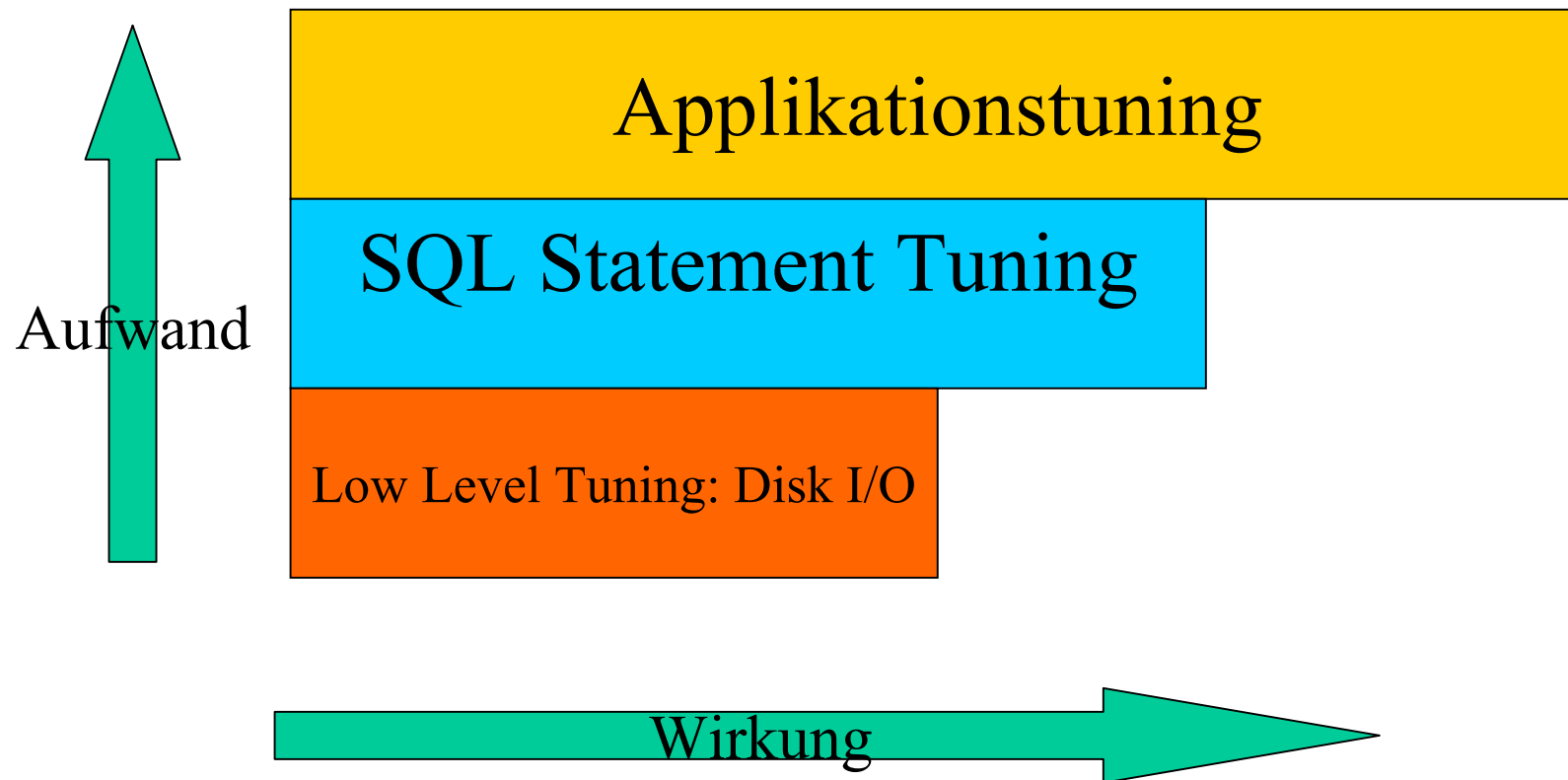


tdisk

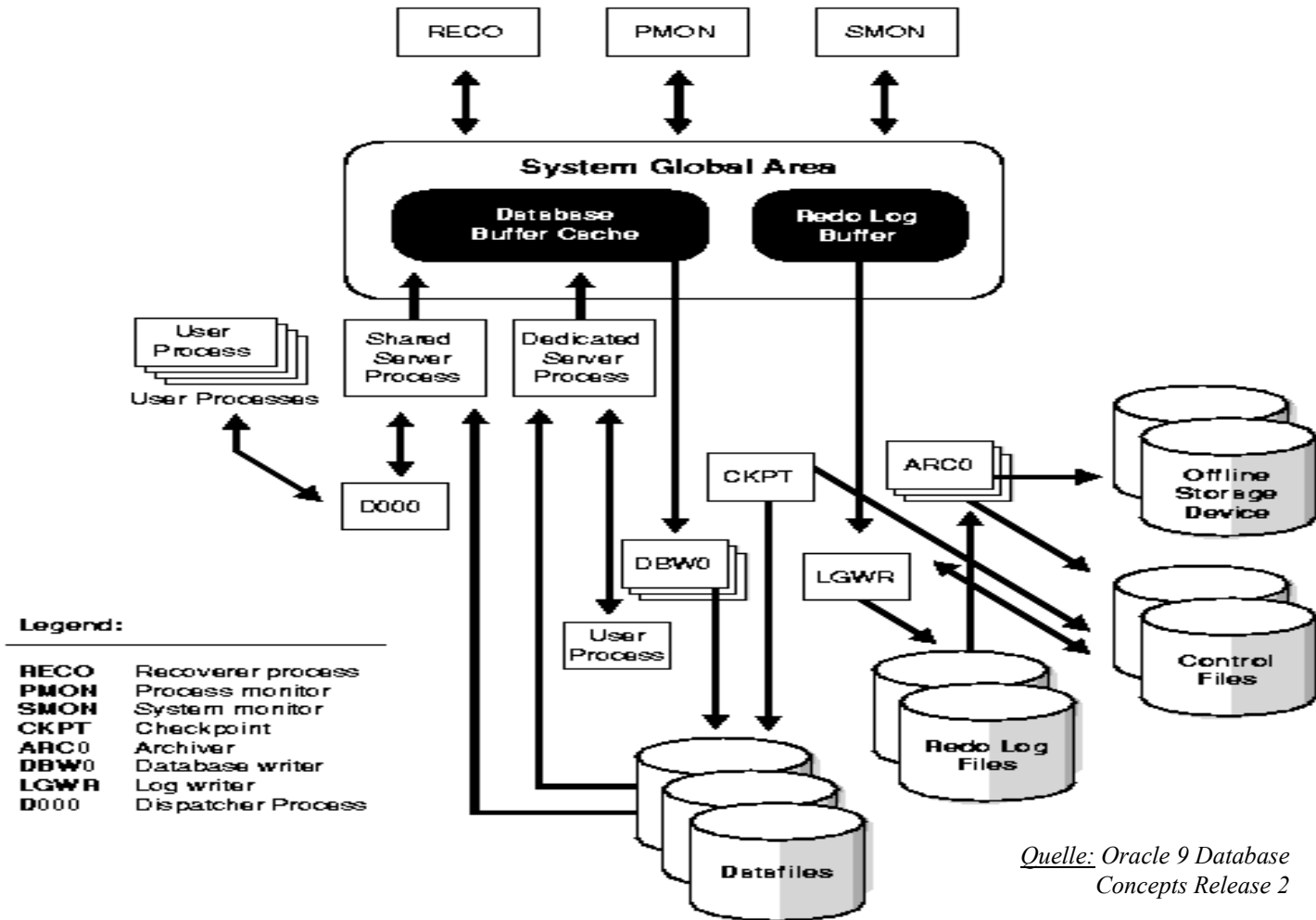
Ein kostenloses Speedometer für Oracle Datenbanken

"Es gibt Lügen, Notlügen und Benchmarks."

Unterschiedliche Tuning Level



Prozessarchitektur



Stichworte

- Stichworte der Veranstaltung:
 - Low Level Tuning
 - Messung von Plattenperformance
 - Asynchrones I/O
 - LogWriter
 - DBWriter
 - ODM

Methode, Idee

- altbewährte Methode:
 - Werkzeugkasten stichprobenartige Untersuchung des Oracle Prozessverhaltens
 - Simulation durch kleines C-Programm
- kleines Testprogramm wurde in C entwickelt, das die elementaren Schreiboperationen simuliert

Themen beispielhaft

- Häufige Fragestellungen
 - Wie schnell sind meine Platten überhaupt?
 - Welchen Performancegewinn bringt der Einsatz von Unix Raw Devices zumindest für die Online Redo Logs?
 - Welcher RAID Level bringt welche Vorteile bzw. Nachteile?
 - Um welchen Faktor kann meine Datenbankapplikation langsamer werden, wenn ich Remote Spiegel zur Datensicherung einschalte?
 - Welche Datenbank-Blockgröße ist am besten für meine Anwendung?

Einsatzmöglichkeiten

Einsatzmöglichkeiten von `tdisk`

- Ermittlung von Abschätzungen über das I/O Limit auf unterschiedlichen Devices
- Vergleichen von unterschiedlichen Devices
- Ermittlung maximaler Schreibgeschwindigkeit von LGWR und DBWR
- Rückwirkung auf Tuning Maßnahmen

Überblick

Features von `tdisk`

- `tdisk` kann von beliebigen Dateien lesen und schreiben: Filesysteme, UFS Filesystemen, Veritas Raw Volumes als auch Raw Devices
- Blockgröße in kByte und Anzahl der zu schreibenden Blöcke können eingestellt werden
- Für das Lesen oder Schreiben kann mit einem Zufallsgenerator auf der Platte positioniert werden
- Herkömmliches synchrones als auch asynchrones I/O sind möglich
- Das Flag `D_SYNC` beim `open ()` Befehl läßt sich setzen

Operating Systeme

- Aufgrund der Portabilität von Oracle sollten die Aussagen prinzipiell auf allen unterstützten Operating Systemen gelten.
- Bisherige Anwendungen und Beispiele setzen den Fokus auf Sun Solaris.

Performance Tools/Solaris

- `iostat` - iteratively reports terminal, disk, and tape I/O and CPU activity
- `vmstat` - report virtual memory statistics
- `mpstat` - report per-processor statistics
- `sar` - system activity reporter
- `lockstat` - report kernel lock and profiling statistics
- `netstat` - show network status
- **proc tools in `/usr/proc/bin`**
- `prstat` (`top`)
- `truss`, `sotrust`, `apptrace`
- `mementool` - unsupported

Anwendung pmap

```
oracle@solaris8:/export/home/oracle>pmap -x 780
780:   oracleOID (DESCRIPTION=(LOCAL=YES) (ADDRESS=(PROTOCOL=beq)))
  Address   Kbytes Resident Shared Private Permissions      Mapped File
08041000     28      28      -      28 read/write/exec   [ stack ]
08050000  21532  10808   6612   4196 read/exec         oracle
09566000    188    188    124    64 read/write/exec   oracle
09595000    248    228      -    228 read/write/exec   [ heap ]
20000000  25016  25016      -   25016 read/write/exec/shared [ ism shmid=0x1 ]
.....
.....
.....
DFBF1000      4      4      -      4 read/write/exec   ld.so.1
-----
total Kb   52744   38988   8572   30416
oracle@solaris8:/export/home/oracle>
```

Anwendung truss

```
1704: open64("/export/home/oracle/db/OID/data/idncat1_ORCLTST.dbf", O_RDONLY) = 12
1704: close(12) = 0
1704: open64("/export/home/oracle/db/OID/data/idncat1_ORCLTST.dbf", O_RDWR|O_DSYNC) = 12
1704: getrlimit64(RLIMIT_NOFILE, 0x08045F80) = 0
1704: fstat64(393, 0x08045EF0) = 0
1704: fstat64(392, 0x08045EF0) Err#9 EBADF
1704: fcntl(12, F_DUP2FD, 0x00000188) = 392
1704: close(12) = 0
1704: fcntl(392, F_SETFD, 0x00000001) = 0
1704: ioctl(392, 0x0403, 0x08045F64) Err#25 ENOTTY
1704: fstatvfs64(392, 0x08045FC8) = 0
1704/1: -> libc:directio(0x188, 0x1)
1704: ioctl(392, 0x2000664C, 0x00000001) = 0
1704/1: <- libc:directio() = 0
1704: fcntl(392, F_GETFL, 0x00000000) = 8258
1704: fstat64(392, 0x08045F9C) = 0
1704: fcntl(392, F_SETLK64, 0x080460BC) = 0
1704: pread64(407, "1502\0\012\0\0\0E501\0\0"..., 2048, 36864) = 2048
1704: pread64(407, "1502\0\0 $\0\0\0A001\0\0"..., 2048, 73728) = 2048
1704: stat64("/export/home/oracle/db/OID/data/icncat1_ORCLTST.dbf", 0x08046034) = 0
1704: open64("/export/home/oracle/db/OID/data/icncat1_ORCLTST.dbf", O_RDONLY) = 12
```

Aufruf von `tdisk`

Kurzbeschreibung des Aufrufes:

```
usage: tdisk [-i input-file] [-o output-file] [-b block-size] [-c count] [-n] [-a] [-m mode] [-l] [-p] [-v]
input-file :   file to test read performance; default: no read test
output-file :  file to test write performance; default: no write test
block-size :   size of data blocks for read/write test, specified in kBytes
count :        number of to blocks to be read/written
-n :           new file, will be created; default: overwrite
-a :           Asynchronous I/O; default: Synchronous I/O
mode :         Flags for File Open, default: O_DSYNC
-l :           llseek, seeking in the file by random
-p :           print CPU time used at the end of the program
-v :           print CPU time used for each read/write call
```

Erweiterungsmöglichkeiten für `tdisk`

- Portierung auf andere Unix Systeme: Anpassung der I/O Calls, ggf. wird der Teil mit Asynch I/O entfernt z.B. unter Linux
- Portierung auf Windows Server Betriebssysteme (SDK oder Public Domain `truss` Utilities)
- Aktivierung von DirectIO
- Performance Vergleich zum Oracle Disk Manager (ODM): Einbau der neuen ODM Calls
- Simulation paralleles Schreiben von mehreren Prozessen (Oracle Parameter `db_writer_processes`, mit dem mehrere DBWR Prozesse gestartet werden)

Suche nach Tools für Windows

- diverse Freeware wie `strace`, teils rudimentär
- Prüfung des Microsoft SDK
- Fremdprodukte
- Process Explorer
- File Monitor

Process Explorer/Windows

The screenshot shows the Process Explorer window with the following data:

Process	PID	CPU	Description	User Name	Priority	Handles	Window...
nvsvc32.exe	580	00		NT-AUTORITÄT\SYSTEM	8	63	
svchost.exe	628	00	Generic Host Process for Win32 Services	NT-AUTORITÄT\SYSTEM	8	159	
wuauclt.exe	1244	00	Client des automatischen Updates von Window...	ATELCO\Andreas Schmidt	8	112	
regsvc.exe	656	00	Remote Registry Service	NT-AUTORITÄT\SYSTEM	8	29	
MSTask.exe	676	00	Taskplaner-Engine	NT-AUTORITÄT\SYSTEM	8	132	
tcpvcs.exe	732	00	TCP/IP Services Application	NT-AUTORITÄT\SYSTEM	8	116	
vmware-authd.ex	760	00		NT-AUTORITÄT\SYSTEM	8	75	
vmnetdhcp.exe	804	00	VMnet DHCP Service	NT-AUTORITÄT\SYSTEM	8	36	
vmnat.exe	836	00		NT-AUTORITÄT\SYSTEM	8	89	
WinMgmt.exe	872	00	Windows-Verwaltungsinstrumentation	NT-AUTORITÄT\SYSTEM	8	120	
ums.exe	936	00		NT-AUTORITÄT\SYSTEM	8	112	
init.exe	1004	00		NT-AUTORITÄT\SYSTEM	8	133	
CPD.EXE	1264	00	McAfee Firewall	NT-AUTORITÄT\SYSTEM	8	142	
CPD.EXE	1260	00	McAfee Firewall	ATELCO\Andreas Schmidt	8	51	
devldr32.exe	1344	00	DevLdr32	ATELCO\Andreas Schmidt	8	68	
ORACLE.EXE	1548	00	Oracle RDBMS Kernel Executable	NT-AUTORITÄT\SYSTEM	8	274	
lsass.exe	264	00	LSA-Exe und Server-DLL	NT-AUTORITÄT\SYSTEM	9	288	
csrss.exe	196	00	Client Server Runtime Process	NT-AUTORITÄT\SYSTEM	13	466	
Explorer.EXE	1284	00	Windows Explorer	ATELCO\Andreas Schmidt	8	435	C:\Progra...
cmd.exe	256	00	Windows NT-Befehlsprozessor	ATELCO\Andreas Schmidt	8	22	DOS Box ...
sqlplus.exe	308	00		ATELCO\Andreas Schmidt	8	58	
procexp.exe	1176	01	Sysinternals Process Explorer	ATELCO\Andreas Schmidt	13	74	Process ...
AHQTB.EXE	1472	00	Creative AudioHQ	ATELCO\Andreas Schmidt	8	52	
realsched.exe	1496	00	RealNetworks Scheduler	ATELCO\Andreas Schmidt	8	57	
CMGrdian.exe	1516	00	McAfee Guardian Agent	ATELCO\Andreas Schmidt	8	75	
MxOALDR.EXE	1524	00	Maxtor MxO Auto Loader Application	ATELCO\Andreas Schmidt	8	39	
OSA.EXE	1528	00	Microsoft Office Wrapper	ATELCO\Andreas Schmidt	8	45	
OneTouch.exe	1536	00	Maxtor OneTouch Detection	ATELCO\Andreas Schmidt	8	38	
internat.exe	1552	00	Sprachanzeigeprogramm	ATELCO\Andreas Schmidt	8	30	
RuLaunch.exe	1572	00	RuLaunch	ATELCO\Andreas Schmidt	8	40	Update M...
RFE32.EXE	1640	00	Programmer's File Editor for Windows NT	ATELCO\Andreas Schmidt	8	34	Program...

Handle	Type	Access	Name
0x2C8	File	0x0012019F	G:\database\DEV\data\TABLESDEV.DBF
0x2CC	File	0x00120089	C:\oracle\ora92\rdms\mesg\oraus.msb
0x2DC	File	0x00120089	C:\oracle\ora92\rdms\mesg\oraus.msb
0x2E4	File	0x0012019F	G:\database\DEV\data\TOOLSDEV.DBF
0x2E8	File	0x0012019F	G:\database\DEV\data\TOOLSDEV.DBF
0x2EC	File	0x0012019F	G:\database\DEV\data\EXAMPLE1.DBF
0x2F0	File	0x0012019F	G:\database\DEV\data\USERSDEV.DBF
0x2F4	File	0x0012019F	G:\database\DEV\data\INDEXESDEV.DBF
0x2F8	File	0x0012019F	G:\database\DEV\data\USERSDEV.DBF
0x2FC	File	0x0012019F	G:\database\DEV\data\TABLESDEV.DBF
0x300	File	0x0012019F	G:\database\DEV\data\TABLESDEV.DBF
0x304	File	0x0012019F	G:\database\DEV\data\INDEXESDEV.DBF
0x308	File	0x0012019F	G:\database\DEV\data\TEMPDEV.DBF
0x30C	File	0x0012019F	G:\database\DEV\data\RBBSDEV.DBF
0x310	File	0x0012019F	G:\database\DEV\data\TEMPDEV.DBF
0x314	File	0x0012019F	G:\database\DEV\data\SYSTEMDEV.DBF
0x318	File	0x0012019F	G:\database\DEV\data\SYSTEMDEV.DBF
0x31C	File	0x0012019F	G:\database\DEV\data\RBBSDEV.DBF
0x320	File	0x0012019F	G:\database\DEV\data\ADDED.DBF
0x324	File	0x0012019F	G:\database\DEV\data\OLAPBAT.DBF
0x328	File	0x0012019F	G:\database\DEV\data\ADDED.DBF

ORACLE.EXE pid: 1548 Refresh Rate: Paused

File Monitor/Windows

The screenshot shows the File Monitor application window with a menu bar (File, Edit, Options, Volumes, Help) and a toolbar. The main area displays a table of file operations. The table has columns for #, Time, Process, Request, Path, Result, and Other. The operations are performed by ORACLE.EXE:1548 on the file G:\DATABASE\DEV\DATA\TABLESDEV.DBF. The requests include IRP_MJ_CREATE, IRP_MJ_READ, IRP_MJ_WRITE, IRP_MJ_CLEANUP, and IRP_MJ_CLOSE. The results are all SUCCESS. The 'Other' column contains various options and offsets.

#	Time	Process	Request	Path	Result	Other
1	20:46:04.780	ORACLE.EXE:1548	IRP_MJ_CREATE	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Options: Open NoBuffer Access: All
2	20:46:04.780	ORACLE.EXE:1548	IRP_MJ_CREATE	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Options: Open NoBuffer Access: All
3	20:46:04.780	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1890304 Length: 2048
4	20:46:04.790	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1892352 Length: 2048
5	20:46:04.790	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1894400 Length: 2048
6	20:46:04.790	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1896448 Length: 2048
7	20:46:04.790	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1898496 Length: 2048
8	20:46:04.800	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1900544 Length: 2048
9	20:46:04.800	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1902592 Length: 2048
10	20:46:04.800	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1904640 Length: 2048
11	20:46:04.800	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1906688 Length: 2048
12	20:46:04.810	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1908736 Length: 2048
13	20:46:04.810	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1910784 Length: 2048
14	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1912832 Length: 2048
15	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1890304 Length: 2048
16	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1892352 Length: 2048
17	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1894400 Length: 2048
18	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1896448 Length: 2048
19	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1898496 Length: 2048
20	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1900544 Length: 2048
21	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1902592 Length: 2048
22	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1904640 Length: 2048
23	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1906688 Length: 2048
24	20:46:05.010	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1908736 Length: 2048
25	20:46:05.020	ORACLE.EXE:1548	IRP_MJ_WRITE	G:\database\dev\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1910784 Length: 2048
26	20:46:05.020	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1914880 Length: 2048
27	20:46:05.080	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1916928 Length: 2048
28	20:46:05.080	ORACLE.EXE:1548	IRP_MJ_READ	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	Offset: 1918976 Length: 2048
29	20:46:08.896	ORACLE.EXE:1548	IRP_MJ_CLEANUP	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	
30	20:46:08.896	ORACLE.EXE:1548	IRP_MJ_CLOSE	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	
31	20:46:08.896	ORACLE.EXE:1548	IRP_MJ_CLEANUP	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	
32	20:46:08.896	ORACLE.EXE:1548	IRP_MJ_CLOSE	G:\DATABASE\DEV\DATA\TABLESDEV.DBF	SUCCESS	

strace/Windows

```
C:\Programme\sysinternals\strace-0.3\app\Release>strace
```

Usage:

```
strace -p <pid>
```

```
strace <cmdline>
```

```
3441 1232 1580 NtFsControlFile (28, 0, 0x0, 0x0, 0x90028, 0x0, 0, 0,
... {status=0x0, info=0}, 0x0, ) == 0x0
3442 1232 1580 NtCreateFile (0x80100080, {24, 0, 0x40, 0, 1224828,
"\??\C:\oracle\ora92\SQLPlus\mesg\spplus.msb"}, 0x0, 128, 1, 1, 2144,
0, 0, ... 148, {status=0x0, info=1}, ) == 0x0
3443 1232 1580 NtSetInformationFile (148, 1224904, 8, Position, ...
{status=0x0, info=0}, ) == 0x0
3444 1232 1580 NtReadFile (148, 0, 0, 0, 256, 0x0, 0, ... {status=0x0,
info=256}, ".....", ) == 0x0
3568* 1232 1580 NtClose (64, ... ) == 0x0
```

„dynamische Mount Optionen“

FILESYSTEMIO_OPTIONS, Parameter type String

Syntax:

FILESYSTEMIO_OPTIONS = {none | setall | directIO | asynch}

Default value: There is no default value.

Parameter class: Dynamic: ALTER SESSION, ALTER SYSTEM

truss zeigt zum Beispiel:

Der Aufruf von `directio()` aus der C Library

generiert wohl intern den folgenden Systemcall:

`ioctl - control device`

Technischer Hintergrund ODM

- Oracle Disk Manager (ODM) mit Version 9 im Produktumfang des Datenbankserver enthalten
- Anstelle der System Calls für I/O (z.B. open, close, read, readv, write, writev, pwrite64, aio_write64) werden eigene Funktionen zur Dateiverwaltung eingesetzt zur:
 - Identifikation von Dateien
 - Datei-Erzeugung und -Management
 - Datei I/O Verarbeitung
- Softwarelieferanten liefern in speziellen Produkten (i.d.R. lizenzpflichtig) passende Bibliotheken mit. Ein symbolischer Link (`$ORACLE_HOME/lib/libodm9.so`) wird auf die jeweilige Library gesetzt.
- Angestrebte Ziele:
 - Verbessertes Management von Datenbankdateien
 - Bessere Performance (Entlastung der Prozesstabellen)
 - Schnellere Dateierzeugung

Einsatzmöglichkeiten für ODM

Aus folgenden Gründen macht es Sinn, anhand einer Testdatenbank den Einsatz von ODM in bestimmten Projekten zu untersuchen:

- Entlastung der Prozeßtabellen mit den vielen I/O Deskriptoren → Reduktion Systemoverhead
- Bei hohem Logwriter-Volumen durch den asynchronen Ablauf (KAIO) schnelleres Schreiben
- ODM ist die technische Basis von RAC. Dies bedeutet, dass die Erkenntnisse über Stabilität und Performanceverbesserungen auch als erste Basis für einen potentiellen RAC Einsatz dienen können.

Schnittstellen der ODM Lib

```
oracle@vmlinux73:~/ora9i2/lib> nm -g libodmd9.so
00001be0 A _DYNAMIC
00001bbc A _GLOBAL_OFFSET_TABLE_
00001c80 A __bss_start
        w __deregister_frame_info@@GLIBC_2.0
        w __gmon_start__
        w __register_frame_info@@GLIBC_2.0
00001c80 A _edata
00001c98 A _end
00000b00 ? _fini
00000788 ? _init
000009e0 T odm_abort
00000a60 T odm_cancel
00000a30 T odm_cleanup
000009d0 T odm_commit
000009c0 T odm_create
000009f0 T odm_delete
00000910 T odm_discover
00000980 T odm_error
00000970 T odm_fini
00000a00 T odm_identify
00000960 T odm_init
00000a40 T odm_io
00000a50 T odm_ioerror
00000a90 T odm_mname
00000a70 T odm_posted
00000a10 T odm_reidentify
00000a80 T odm_resize
00000a20 T odm_unidentify
        U strncpy@@GLIBC_2.0
```

weitere Infos etc.

- [<http://developers.sun.com/solaris/articles/solUFSdiskIO.html>] zeigt Unterschiede von `fdatasync()` and `directIO()` system calls bezogen auf SCSI und IDE disks
- [Jim Mauro, Richard McDougall: Solaris Internals] beschreibt das Design des Betriebssystems Solaris8 mit vielen Details
- Oracle Papiere zu ODM
- Unix Man Pages
- und ... und ... und ...