



Troubleshooting and Tuning Oracle RAC on Linux

Edward Whalen
Performance Tuning
Corporation



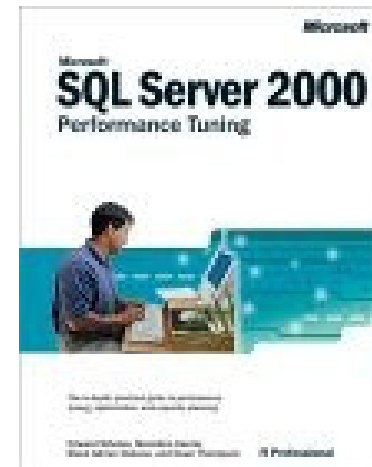
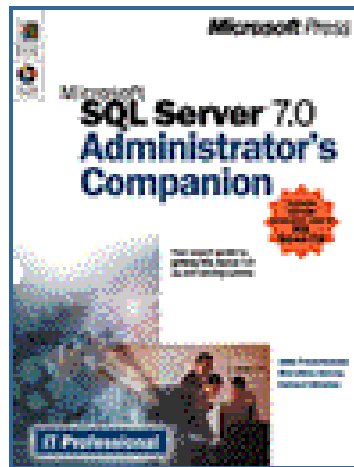
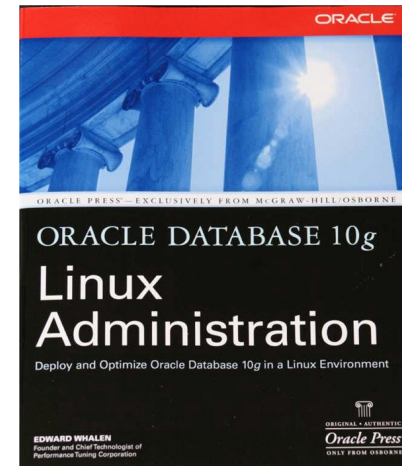
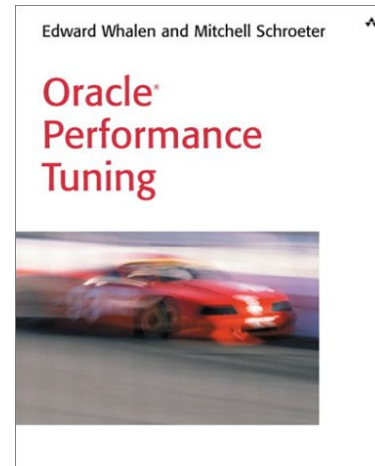
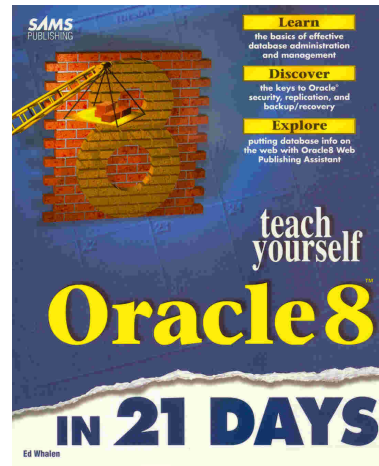
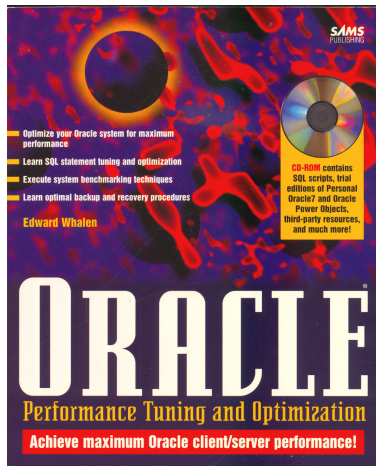
Introduction

Edward Whalen

- Founder and CTO of Performance Tuning® Corporation with more than 20 years technology experience.
- Oracle and SQL Server Database Performance Expert.
- Author of 8 books and various technical papers on Oracle and SQL Server database technologies.
- Former UNIX System Developer

Introduction

Edward Whalen's Books





Presentation Overview

- Brief Overview of Linux
- Overview of the Oracle Database Server on Linux
- Introduction to Oracle Real Application Clusters (RAC) on Linux
- 32-bit vs. 64-bit Linux
- Troubleshooting Oracle RAC on Linux
- Tuning Oracle RAC on Linux



Brief Overview of Linux

- Linux is based on a modular non-microkernel architecture
 - All device drivers share the same memory with the kernel
 - Device drivers are modular and loadable/unloadable
- Linux is multi-tasking
- Linux is a virtual-memory OS
- Most kernel parameters can be modified on the fly



Brief Overview of Linux

- 1991 – Linus Torvalds led the development of Linux at the University of Helsinki in Finland where the first version was released.
- 1999 – Linux version 2.2 kernel released
- 2001 – Linux version 2.4 kernel released
- Late 2003 – Linux version 2.6 kernel released



Linux 2.6 Kernel Features

- Support for NUMA Systems
 - Non-Uniform Memory Access
- Support for Hyperthreading
- Support for larger systems and more devices
 - Limit of 255 major devices has grown to 4096 major devices
- I/O Subsystem Improvements
- Improved Networking (including NFS)
- Better support for LVM
- 32-bit and 64-bit support



Oracle on Linux

- Oracle announced support for Linux in 1998 and has been behind it ever since
- Oracle supports most products on Linux
 - Database Server and RAC
 - Oracle Application Server (32-bit mode)
 - Oracle Developer Suite
 - Oracle E-Business Suite (32-bit)



Intro to Oracle RAC on Linux

- Oracle RAC on Linux has been around since RAC was released
 - Oracle9i
- Oracle Database 9i and Oracle Database 10g both available in x86 and x64 versions
- Support for Ethernet or Infiniband Interconnect



32-bit vs. 64-bit Linux and Oracle

- 64-bit Linux
 - X64 (AMD Opteron and Intel EM64T)
 - Intel Itanium2

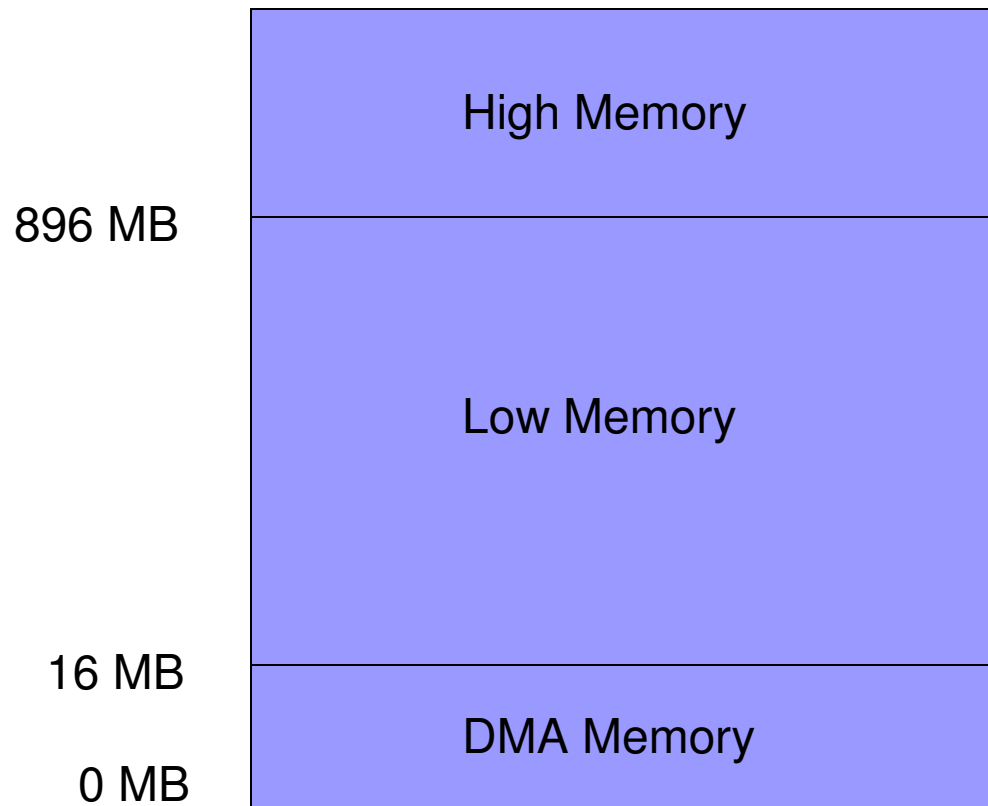


x86 vs x64 Architectures

- x86 – 32-bit
 - Virtual Memory 4 GB
 - Physical Memory 4 GB
 - Physical Memory with PAE 64 GB
- X64 – 64-bit
 - Virtual Memory 256 Terabytes (architecture supports 16 exabytes)
 - Physical Memory 1 Terabyte (later will be increased to 4 petabytes)

32-bit Linux Memory

- Memory is divided into 3 zones





64-bit Advantages

- No PAE (Page Address Extension) to worry about
- Everything is Low Memory
- Large SGAs supported without Indirect Data Buffers
- 64-bit hardware is available today



Troubleshooting

- Attitude and methodology
- Finding the Logs
- Areas of Focus (Common Problems)



Troubleshooting Attitude

- Troubleshooting
 - 50% Skill
 - 50% Experience
 - 50% Attitude
- Keep Positive
- Don't give up
- Stay Focused



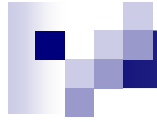
Finding the Logs

- Oracle Alert Log
 - Database Instance
 - ASM Instance
- Cluster Logs
 - 10gR1
 - \$ORA_CRS_HOME/<service>/logs
 - 10gR2
 - \$ORA_CRS_HOME/logs/<service>



Areas of Focus

- Interconnect
 - Crucial for RAC performance
- Shared Disk
 - Performance
 - Stability
- CPU
 - Cluster Services are key



Tuning

- Monitoring Oracle RAC on Linux
- Tuning Oracle RAC on Linux



Monitoring RAC on Linux

- Linux Monitoring
- Oracle Monitoring



Linux Monitoring

- top
- sar
- free
- iostat
- vmstat



Top

- Pros

- Dynamic
- Provides good instantaneous information

- Cons

- Doesn't log

Top

```
top - 11:46:50 up 5 days, 12:38, 1 user, load average: 0.24, 0.14, 0.13
Tasks: 166 total, 1 running, 165 sleeping, 0 stopped, 0 zombie
Cpu0  :  1.0% us,  1.7% sy,  0.0% ni, 95.7% id,  1.7% wa,  0.0% hi,  0.0% si
Cpu1  :  2.0% us,  1.7% sy,  0.0% ni, 96.3% id,  0.0% wa,  0.0% hi,  0.0% si
Mem:   2053592k total, 2017852k used,  35740k free,  213340k buffers
Swap:  8388600k total,  73208k used,  8315392k free, 1189184k cached
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
3354	root	15	0	21224	8660	1588	S	0.3	0.4	3:45.92	hald
3483	root	16	0	6560	1272	948	S	0.3	0.1	11:38.76	/bin/sh
/etc/init.d											
12136	root	16	0	6280	1060	756	R	0.3	0.1	0:00.05	top c
1	root	16	0	4752	596	496	S	0.0	0.0	0:05.08	init [5]
2	root	RT	0	0	0	0	S	0.0	0.0	0:25.29	[migration/0]
3	root	34	19	0	0	0	S	0.0	0.0	0:02.51	[ksoftirqd/0]
4	root	RT	0	0	0	0	S	0.0	0.0	0:22.97	[migration/1]
5	root	34	19	0	0	0	S	0.0	0.0	0:01.11	[ksoftirqd/1]
6	root	5	-10	0	0	0	S	0.0	0.0	0:00.02	[events/0]
7	root	5	-10	0	0	0	S	0.0	0.0	0:00.01	[events/1]
8	root	8	-10	0	0	0	S	0.0	0.0	0:00.00	[khelper]
9	root	15	-10	0	0	0	S	0.0	0.0	0:00.00	[kacpid]
45	root	5	-10	0	0	0	S	0.0	0.0	0:00.00	[kblockd/0]
46	root	5	-10	0	0	0	S	0.0	0.0	0:00.00	[kblockd/1]
47	root	15	0	0	0	0	S	0.0	0.0	0:00.06	[khubd]
59	root	15	0	0	0	0	S	0.0	0.0	0:06.93	[pdflush]



Sar

- Long term statistics
- Automatically gathered
- Many different counters
 - CPU
 - sar
 - Network
 - sar -n DEV
 - IO
 - Paging
 - Etc.

Sar

■ sar

Linux 2.6.9-34.ELlargesmp (ptcl.perftuning.com) 10/15/2006

12:00:01 AM	CPU	%user	%nice	%system	%iowait	%idle
12:10:01 AM	all	2.93	0.00	2.40	1.01	93.66
12:20:01 AM	all	2.80	0.00	2.34	1.08	93.79
12:30:01 AM	all	2.81	0.00	2.38	1.12	93.69
12:40:01 AM	all	2.85	0.00	2.41	1.03	93.71
12:50:01 AM	all	2.84	0.00	2.40	1.05	93.72
01:00:01 AM	all	2.93	0.00	2.39	1.04	93.64
01:10:01 AM	all	2.82	0.00	2.37	1.13	93.68

■ sar -n DEV

Linux 2.6.9-34.ELlargesmp (ptcl.perftuning.com) 10/15/2006

12:00:01 AM	IFACE	rxpck/s	txpck/s	rxbyt/s	txbyt/s	rxcmp/s	txcmp/s	rxmcsst/s
12:10:01 AM	lo	6.42	6.42	1256.12	1256.12	0.00	0.00	0.00
12:10:01 AM	eth0	0.93	0.85	120.51	133.71	0.00	0.00	0.00
12:10:01 AM	eth1	17.72	15.08	12128.13	8669.81	0.00	0.00	0.00
12:10:01 AM	sit0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12:20:01 AM	lo	6.24	6.24	1199.23	1199.23	0.00	0.00	0.00
12:20:01 AM	eth0	1.16	1.28	147.09	767.49	0.00	0.00	0.00
12:20:01 AM	eth1	7.95	6.45	3584.42	2146.04	0.00	0.00	0.00
12:20:01 AM	sit0	0.00	0.00	0.00	0.00	0.00	0.00	0.00



Free


- Displays memory statistics
- Displays low and high memory
- Can be used to strip off cached memory



Free (32-bit)

```
[root@pg04 ~]# free -lm
```

	total	used	free	shared	buffers	cached
Mem:	1519	1183	335	0	3	699
Low:	879	545	334			
High:	639	638	1			
-/+ buffers/cache:		480	1038			
Swap:	1023	76	947			



Free (64-bit)

```
[root@ptcl ~]# free -lm
```

	total	used	free	shared	buffers	cached
Mem:	2005	1971	34	0	208	1161
Low:	2005	1971	34			
High:	0	0	0			
-/+ buffers/cache:		601	1404			
Swap:	8191	71	8120			



lostat

- Used to monitor I/O
- Provides information on
 - IOPS
 - Queue depths
 - Queue time
 - Response time



lstat

```
avg-cpu:  %user   %nice    %sys %iowait  %idle
           1.45    0.00    1.70   0.70   96.15
```

```
Device:   rrqm/s  wrqm/s   r/s    w/s  rsec/s  wsec/s   rkB/s   kB/s  avgrq-sz  avgqu-sz   await  svctm   %util
sda       0.00    7.80    0.00   9.80   0.00  140.80    0.00   70.40   14.37    0.02    1.97   0.04   0.04
sdb       0.00    0.00    3.00   1.80   36.00   17.00   18.00    8.50   11.04    0.02    3.56   3.25   1.56
sdc       0.00    0.00    0.00   0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00   0.00   0.00
dm-0     0.00    0.00    0.00   4.80    0.00   38.40    0.00   19.20    8.00    0.00    0.00   0.00   0.00
dm-1     0.00    0.00    0.00   0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00   0.00   0.00
dm-2     0.00    0.00    0.00   9.90    0.00   79.20    0.00   39.60    8.00    0.02    2.41   0.04   0.04
dm-3     0.00    0.00    0.00   0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00   0.00   0.00
dm-4     0.00    0.00    0.00   2.20    0.00   17.60    0.00    8.80    8.00    0.00    0.82   0.09   0.02
dm-5     0.00    0.00    0.00   0.70    0.00    5.60    0.00    2.80    8.00    0.00    2.57   0.43   0.03
dm-6     0.00    0.00    0.00   0.00    0.00    0.00    0.00    0.00    0.00    0.00    0.00   0.00   0.00
```

VMStat

- Provides information on system virtual memory
 - Memory
 - Paging

```
■ [root@ptcl ~]# vmstat
■ procs -----memory----- ---swap-- -----io----- --system-- ----cpu----
■  r  b   swpd   free   buff  cache   si   so    bi    bo   in    cs  us  sy  id  wa
■  0  0   73196  36140 213592 1188396    0    0    14    31   29    38  3  2  94  1
```



Oracle Monitoring

- AWR
- Dynamic Performance Views



Tuning

- Tuning the hardware (A.K.A. sizing)
- Tuning Linux
- Tuning the Oracle Database Server



Tuning the Hardware

- I/O subsystem
- Interconnect
- Memory
 - 32-bit vs. 64-bit
 - Standard Oracle Database memory tuning



Tuning Linux I/O

- Sizing
- I/O schedulers (2.6 kernel)
- Network tuning
- Oracle Database and Oracle RAC process priorities



I/O Sizing

- It's all about the number of disk drives
 - 125 IOPS per disk
- Avoid slow disk drives (SATA)
- Avoid RAID 5 for Oracle data files



Automatic Storage Management (ASM)

- Automatic Storage Management
- Use external redundancy with ASM
 - We have found 4-drive RAID 1/0 LUNS striped with ASM works well
- Adding disks to diskgroups causes high overhead due to re-striping
 - Do this during off-hours



I/O Schedulers

- Linux 2.6 Kernel supports four schedulers
 - Completely Fair Queuing
 - Deadline
 - NOOP
 - Anticipatory
- Defined in the grub boot file



Completely Fair Queuing

- CFQ
- Default in 2.6 kernel
- Maintains per process I/O queues
- Provides generally well balanced I/Os
- Defined by:
 - `elevator=cfg`



Deadline

- Optimized for short latency
- Uses round-robin for I/O distribution
- Avoids I/O starvation that can occur with other schedulers
- Defined by:
 - elevator=deadline



NOOP

- Uses a FIFO
- Achieves minimal amount of instructions per I/O
- Assumes optimization will occur at the device
- Optimal for SAN/NAS devices
- Defined by:
 - `elevator=noop`



Anticipatory Elevator

- Similar to elvtune in 2.4 kernel
- Introduces an artificial delay
- Allows driver to perform aggressive elevator sorting
- For small or slow I/O subsystems
- Trades latency for throughput
- Defined by:
 - elevator=as



The RAC Interconnect

- Critical to Oracle RAC performance
 - Used for lock and state traffic
 - Used for passing data blocks (Cache Fusion)
- Must be robust
- Must be fast
- Performs optimally if dedicated (not shared)



Network Tuning

- Interconnect is crucial to Oracle RAC performance (just like I/O)
- Separate interconnect from public network
- Separate NAS network from interconnect and public network
- Use TCP Offload Engine (TOE) if possible
 - Note: Most latest generation systems include this.
 - Note: iSCSI offload engines are also available



Network Tuning

- Tune the UDP Buffers

- net.core.rmem_max = 2097152
 - net.core.rmem_default = 262144
 - net.core.wmem_default = 262144
 - net.core.wmem_max = 262144

- More is NOT better



Process Priorities

- Sometimes the cluster processes can get starved for CPU cycles
- If so, increase their priorities

```
ps -elf | grep lms
```

```
renice -20 -p pid1, pid2, pid3
```

```
ps -elf | grep ocssd
```

```
renice -20 -p pid
```



Tuning the Oracle Database

- Instance tuning
- Wait based tuning
- Application tuning



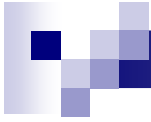
Holistic Tuning

- All layers must be considered
- Interaction is important



Conclusions

- Oracle RAC is tuned much like any other systems
 - Sizing is important
 - The interconnect is crucial
 - Holistic view - analyze the entire system



Q & A

QUESTIONS
ANSWERS



Contact Information

Edward Whalen

CTO

Performance Tuning Corporation

www.perftuning.com

(800) 887-4513

ewhalen@perftuning.com

Performance Tuning® Corporation specializes in assisting IT Departments in resolving these and other Database issues.

Other Appearances

**Meet the Experts Panel
Linux**

OTN Lounge

Tuesday 2:30 – 3:00

Book Signing

OpenWorld Book Store

Thursday 10:00 – 10:30

