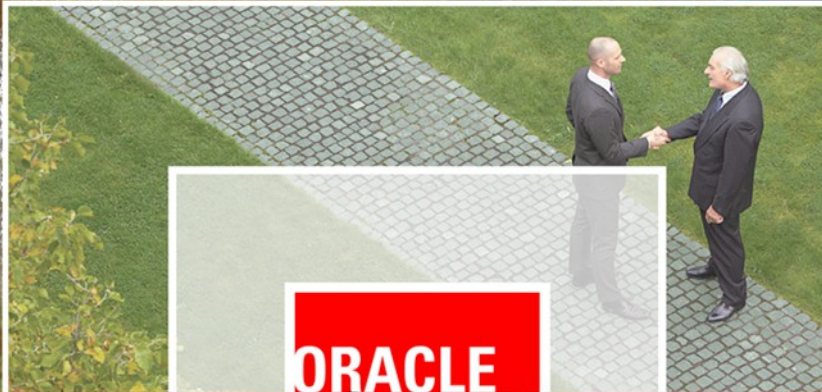


ORACLE®



ORACLE
OPEN
WORLD

Your. Open. World.

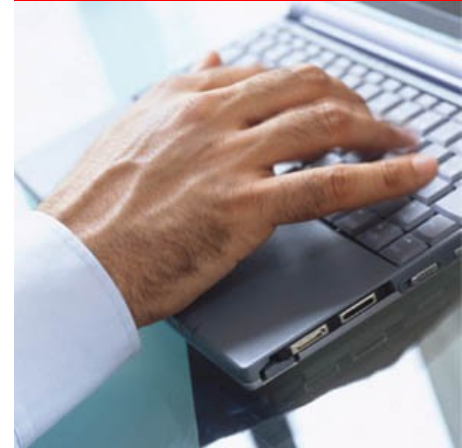
Bulletproof your Linux Systems

Linux and OVM Development

ORACLE®

Bulletproof Linux Agenda

- Demystifying the OS
- Kernel Tuning
- Using Linux's console
- Digging deeper
- Virtualization



Goals of this talk

- Improve your readiness in the case of a crash
- Improve your fluency (reduce opacity) of Linux
- Introduce you to tools and data sources
- Help you to get more performance and ROI

Linux and VM Diagnostics

 Oracle Differentiator

CONFIGURATION

- Capacity planning
- Kernel tuning – generic parameters
- Kernel tuning – workload specific parameters
- *Validated Configurations*




MONITORING

- Monitoring tools
- Attached console
- Debugging for brownouts
- Tools for virtualization debugging
- Ready to file a bug report the first time



ANALYSIS

- One call for support 
- Reading kernel statistics
- Discovering common problems
- Improved performance and ROI, faster time to market

ORACLE®
TRADITIONAL IT

ORACLE®
DEVELOPMENT

Running on... Unbreakable

84,000+ Inte

7,000 Network |

Linux

ORACLE®

External

| 42,000 S's



ORACLE®
ON DEMAND

ORACLE®
UNIVERSITY

ORACLE®

ORACLE®
TRADITIONAL IT

ORACLE®
DEVELOPMENT

Running on... Unbreakable

84,000+ Inte

7,000 Network |

Linux

ORACLE®

External

| 42,000 S's



ORACLE®
ON DEMAND

ORACLE®
UNIVERSITY

ORACLE®

Oracle VM

Oracle tested and supported Server Virtualization



- Enhanced/optimized Xen technology
- 3X more efficient
- Free to download, use and distribute
- Enterprise-quality support
- Templates for faster deployment

Supported and certified with Oracle products.

Bulletproof Linux:

Demystifying the OS



Parts of the OS

- Memory
- IO and filesystems
- Network
- processes and scheduling
- the hypervisor

•Kernel Tuning

- Too many ports of EL to explore this fully.
- Overview of general concepts instead.
- Most important tuning is **Capacity Planning**

What can we tune?

- Anything!
- Is that really a good idea?
- Validated Configurations

Types of parameters

- Running system:
 - `kernel.panic_on_oops = 1`
 - `kernel.panic = 0` (without remote console)
 - `kernel.sysrq = 1`
- Boot time:
 - `nmi_watchdog = timeout`

Validated Configurations

- <http://linux.oracle.com/validated-configurations>
- Most important:
 - kernel parameters
 - combination of product release notes and best practices
- [IMAGE]

IO and Filesystems

- Depends on the FS
- Quick NFS intro
- lvm, dm, dm-snapshot
- IO Elevator - CFQ

Network

- receive and send buffers size
- tcpdump
- arp
- nfsstat
- tcp.rmem_max, tcp.wmem_max

Processes

- Don't schedule in Real Time
- Don't be afraid of renice
- strace -P

Memory

- Most common symptom
- Out of Memory conditions
 - Per process
 - Per physical memory
 - Per address space
- Swap
- Tuning parameters
- shrink_zone
- runaway process

Memory Statistics on Linux

- cached
- free
- swap: *use* is OK, *i/o* is bad
- 0 free memory vs. all cached memory
- hugepages shm (invisible)
- regular shm (cached)
- kswapd finds no file pages and starts to swap process pages (hugepages)
- but you shouldn't be swapping anyway!
- recommendations for shm sizes and swap sizes are BOGUS

vmstat

- numbers for swap
- other numbers
- swappiness (shouldn't be swapping anyway)

Kernel MM behavior

- Kernel tuning is dangerous.
- [IMAGE]
- vm.swappiness
 - increasing swappiness is OK. decreasing swappiness is bad.
 - This is more a Capacity Planning issue
- vm.panic_on_oom

32-bit memory

- the 32-bit address space
- DIAGRAM – LOW MEM
- advantages?
- no difference between 32- and 64-bit processors any more.
 - There is a difference at the OS/Application level.
- Limited process address space
- Limited kernel memory
- lower_zone_protection
 - prevent userspace applications from impinging on kernel memory
- the -hugemem kernel
 - el4 only. Allows 4 gb for the kernel
- the -pae kernel



Hugepages

Bottom line

- Find a validated configuration similar to yours and follow it.
- Keep up to date with the kernels on ULN.

Bulletproof Linux: Preparedness



Getting information out of a running system

- Diagnostic tools
- The best tools are the least obtrusive
- You know the frustration of not having the right tools set up and recording data
- Always have diagnostic tools monitoring your system!

Monitoring tools

- oswatcher / collectl
- kdump/kexec
- crash
- RDA

Oswatcher, in brief

- Time-based system statistics
- Must be taken during the crash or brownout.
- Captures
 - ps
 - slabinfo
 - meminfo
 - mpstat
 - top
 - vmstat
 - iostat
- Comparison with SAR

What can go wrong?

- Diagnosing system slowdowns
- Diagnosing kernel or device driver bugs
- Diagnosing hardware errors
- Three kinds of kernel halts
 - OOPS – code bug
 - BUG – assert statement
 - panic() -- default failure mode
- Same logging required for each error type



Crash readiness

- Linux provides the tools to comprehensively diagnose outages
- **Do not** reboot your system without capturing data.

Bulletproof Linux:

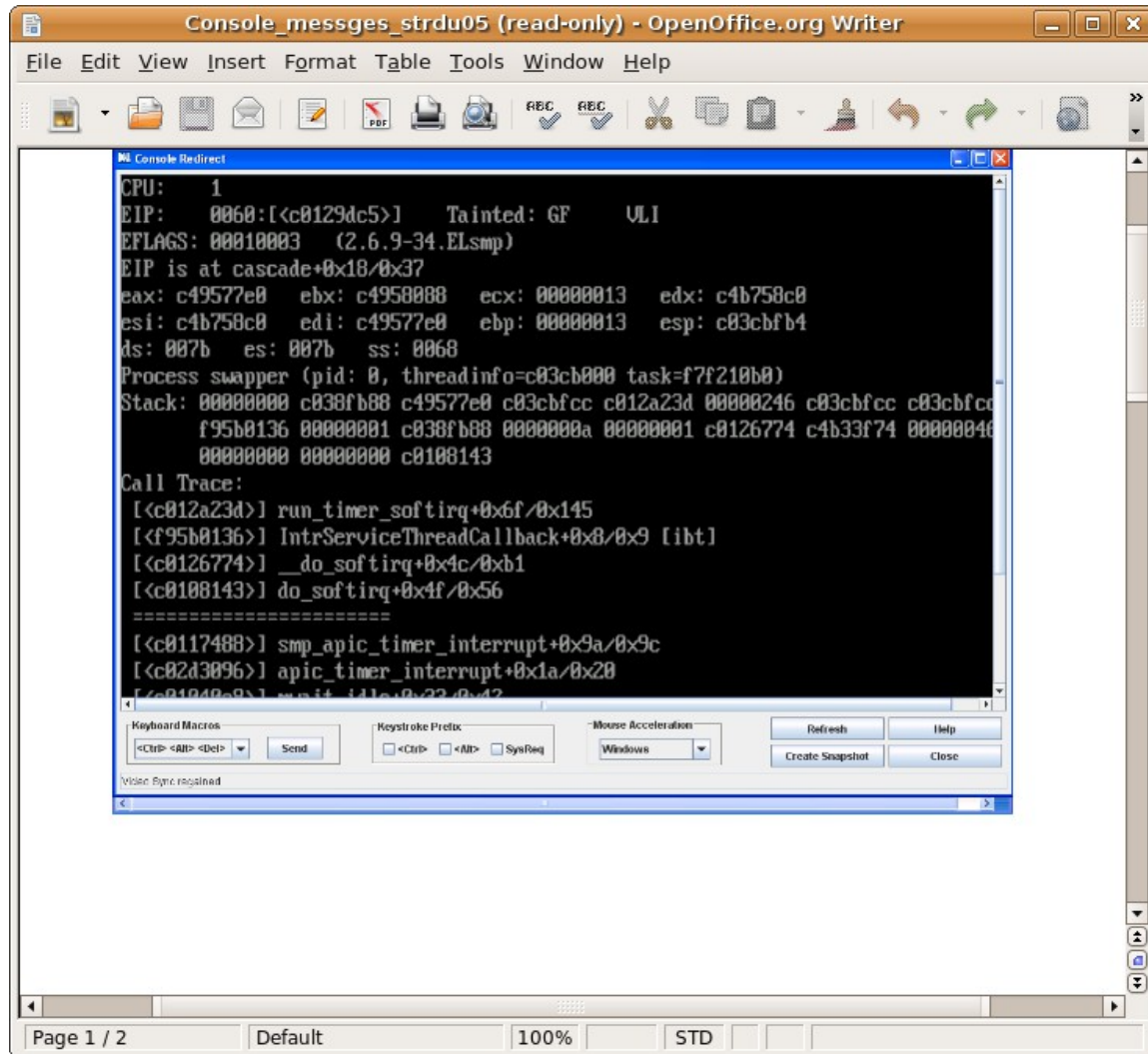
The Console



The Console

- What is it?
- Live heartbeat into the OS
- See the functionality of the OS
- See the “tombstone” messages
- Proven technology
- Plan ahead
 - capture console data
 - need complete dump data

Remote Console Access



Serial Console

- Still the one.
- Been around since the 70s
- Lightwave concentrators
- Expensive to deploy
- The best solution is the lowest-tech: the serial console.
 - Not a critical path for your system
- Disadvantages
 - Hard to find
 - Hard to set up, expensive to run extra cable!
 - Requires another system

Netconsole

```
# modprobe netconsole netconsole=4444@10.0.2.1/eth0,6666@10.0.2.2/00:05:5D:34:11:AF
```

- Formely netdump
- useful for network logging
- not always reliable!
- Console log and vmcore via netdump (EL4)
 - Network cables already connected!
 - Requires a separate netdump server
 - Can be hampered if the networking layer crashes
 - Easy to set up via /etc/init.d/netdump propagate
 - Helpful even when the system is just “slow” or unresponsive
- netconsole (EL4 and EL5)
 - message logging only
 - boot option for netconsole:

```
grub.conf: kernel /vmlinuz-EL5-smp ro root=LABEL=/ console=ttyS0,115200 console=tty debug
```

System core files

- vmcores
- huge
- can get logs
- can get other stuff
- really easy to get from virtualized environments

Dumping Core



- When console messages aren't enough
 - Core files are extremely large
 - Core files are not always required to debug problems
 - Only for the most serious problems
- Why not dump core all the time?
 - HUGE disk space
 - HUGE time to transfer that much memory
 - Only records for catastrophic failures

Kexec

- Soft reboot
- Not proven technology . . . yet
- No data while running – not useful for brownouts
- Could be really neat in the future
- Special crash kernel for postmortem debugging
- Available on EL4 and EL5
 - Only available after a catastrophic failure
 - These do not log console
 - Only capable of extracting the vmcore from the system
 - Separate post-processing required to extract messages
 - Do not necessarily require a second server
 - kdump can dump across the network
 - diskdump requires a crash partition on the root device

Bulletproof Linux:

Digging Deeper





The Kernel Stack

- sysrq-t



Kernel Memory info

- sysrq-m

Out of Memory

- process out of memory
- address space out of memory
- physical out of memory
- not always graceful
- panic_on_oom

Finding runaway processes

- RSS vs VSS
- ulimit hacks
- PGA size
- Things which map the SGA show up huge
 - ok, so a little Oracle specific here. but other processes use shared mem too!

Panic messages

- finding the culprit
- looking for tainted modules
- all about third party modules

Reading kernel messages

- It's all there, you just have to read it
- addresses indicate modules vs kernel
- stacks in reverse order
- not always pure

Bulletproof Linux: Virtualization





It just got easier!

- Cores from running systems
- Allows “live” debug of problems

Common VM problems

- Sizing expectations
- pagecache
- dom-0 vs dom-U
- HVM vs PV
- pausing and cloning
- templates
- make sure you have space in your dest. directory
- capture xm guest console output



Filing a good bug

- OS Version
- rpms installed

Recap

Set up kernel parameters
Have logging technologies
Capacity Planning