Troubleshooting VMware System Problems III

Krishna Raj Raja
VMware

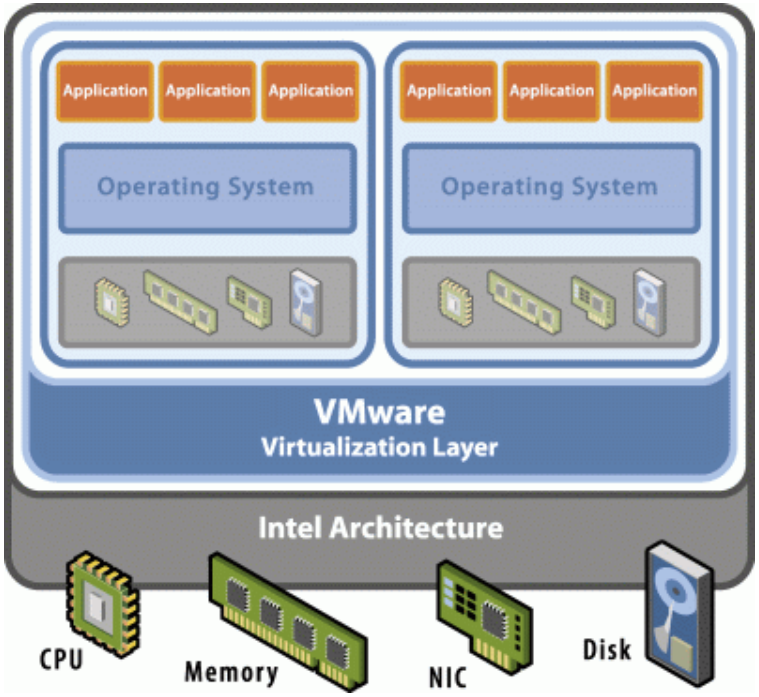
This presentation may contain VMware confidential information.

Agenda

- Problem isolation
- Log files
- Proc nodes
- Performance issues

Problem Isolation: Challenges

- Problem symptoms could be misleading.
- Guest OS/Application/Hardware issues often appear as Virtualization issue
- Non reproducible issues are complicated to isolate.
- Where to begin ?



- **Application**
- **Guest OS**
- **Emulation Layer**
  - **Vlance, vmxnet, LSILogic, Buslogic**
- **Virtualization Layer**
  - **VMkernel, VMM**
- **Hardware**
  - **Memory, CPU, SCSI, FC cables, etc.**
- **External**
  - **SAN/network topology, routers, switches, fabric**

Problem Isolation

Reproducible Issue	Non reproducible Issue
▪Simplify the issue. ▪See if the issue could be isolated to a particular layer ▪Introduce one change at a time until the problem goes away ▪Confirm resolution by repeating the steps ▪Report the problem	▪Record timestamp of every occurrence ▪Collect vm-support dump as after you hit the problem ▪Track every change made to the ESX Server and to the external environment. ▪Offline analysis: Track events in the logs and co-relate to the problem symptoms

The Problem Probability Matrix

Problem does not happen on a different	Application Issue/ configuration	Guest OS Issue/ configuration	Hardware problem	VMware issue
Application	*			*
Guest OS	*	*		
Virtual Machine		*		
ESX Server			*	*
Native Machine		*		*
Virtual Hardware				*
Physical Hardware			*	*

Note: This table does not guarantee problem isolation. Use this as a starting point for troubleshooting

Data Collection

Layer	System State information	Log
Guest/Application	Msinfo32/winmsd Proc nodes for Linux guest	Event logs, System logs, Application logs
Virtual Machine	/proc/vmware/vm/wid/	vmware.log
Console OS	/proc/*	/var/log/messages dmesg /var/log/vmware*/*
Virtualization	/proc/vmware/*	/var/log/vmkernel
Hardware	/proc/net/* /proc/scsi/* /proc/vmware/sched	Management agent logs *hplog –v* *Omreport system [alertlog|cmdlog|esmlog|postlog]* *Omreport chassis [temps|fans|memory]*
External	Storage Array profile, Switch profile	Storage Array logs, FC-Switch, Ethernet Switch logs

Log Files vs. Proc Nodes

- Log files
  - Use "*tail –f*" to monitor the logs live
  - Identify hardware problems and changes in the external environment
  - Logs are rotated, evidences are lost if investigation is delayed
    - Run vm-support and record timestamp as soon as you hit the problem
- Proc Nodes
  - Live system state information, created on demand
  - Configuration differences
    - Intended vs. Real
    - Between Servers
  - Performance monitoring
    - "*watch –d*" to monitor the nodes live
  - Take Proc Node Snapshots at frequent intervals
    - *vm-support -s -i <sleep interval> -d <duration>*
  - /proc/vmware is deprecated in ESX Server 3.0 !

VMkernel Log: Virtual Machine Events

Virtual machine power on

Jun 24 15:52:20 krraja-dev vmkernel: **3:22:21:20.402** cpu0)World: **vm 143**: 894: Successfully created new world: **'vmm0:RHES**

World Name (/proc/vmware/vm /143/names)

VMkernel Uptime (/proc/vmware/uptime) dd:hh:mm:ss:uuu

World id (/proc/vmware/ sched/cpu)

Virtual machine power off

Jun 24 15:52:22 krraja-dev vmkernel: 3:22:21:22.229 cpu0)Host: vm 143: 4884: **destroying world from host**

VMkernel Log: Virtual Machine Events

Virtual machine migration out

Aug 2 16:52:08 krraja-dev vmkernel: 112:00:58:42.610 cpu0)Migrate: vm 200: 3939: Setting migration info ts = 1966929288, src ip = <10.200.0.3> dest ip = <10.200.0.5> Dest wid = 151

Source VMotion IP (/proc/vmware/net/tcpip/config)

Dest migration host IP

Wid on the destination host

Virtual machine migration in

New virtual machine id on the localhost

Sep 13 16:19:27 krraja-dev vmkernel: 154:00:25:12.458 cpu0)Migrate: vm 244: 3939: Setting migration info ts = 4246360415, src ip = <10.200.0.5> dest ip = <0.0.0.0> Dest wid = -1

Heap memory

/proc/vmware/mem

141:00:12:19.860 cpu0:127) WARNING: MemSched: vm 367: 4836: insufficient heap:avail=505K, need=1024K

141:00:12:19.860 cpu0:127) WARNING: Alloc: vm 367: 2731: insufficient memory: unable to admit

VMkernel Log: H/W Events

NMI: Error

Jun 20 10:21:30 liware06 vmkernel: 16:09:07:09.722 cpu0)ALERT: APIC: 1150: Lint1 interrupt on pcpu 0 (port x61 contains 0xa1)

MPS table mode settings (HP/Compaq Systems) – KB 1081

May 11 22:36:30 esx101 vmkernel: 0:00:00:00.00 ALERT: Chipset: 303: no PCI entries - Check BIOS Settings

VMM faults: Virtual machine crash

Apr 21 10:54:59 rws-spb21 vmkernel: 83:03:00:53.673 cpu2:157) WARNING: World: vm 157: 4076: vmm1:VMTest:VMM fault: regs=0x2b10, exc=14, eip=0x5eda5, addr=0xffce75c0

VMkernel Log: Storage Events..

FC - Rescan

vmkernel: 12:20:35:47.618 cpu0)<6>dpc(0): qla2x00: RESCAN .
vmkernel: 12:20:35:47.635 cpu0)<6>dpc(0): qla2x00: RESCAN... done.
vmkernel: 12:20:35:48.452 cpu3)SCSI: 8979: Finished rescan of adapter vmhba2
vmkernel: 12:20:35:48.480 cpu3)SCSI: 8891: Disks have been added or removed from the system.

FC – Link Status

vmkernel: 0:00:21:01.647 cpu2:129) <6>scsi(0): LOOP DOWN detected.
vmkernel: 0:00:21:06.134 cpu0:139) <6>qla2x00: Performing ISP error recovery - ha= 0x12ab394.
vmkernel: 0:00:21:26.285 cpu0:139) <6>scsi(0): Cable is unplugged...
vmkernel: 0:00:21:53.852 cpu2:129) <6>scsi(0): LIP reset occurred.
vmkernel: 0:00:21:53.913 cpu2:129) <6>scsi(0): LIP reset occurred.
vmkernel: 0:00:21:54.032 cpu2:129) <6>scsi(0): LOOP UP detected.

lpfc0:1303:LKe:Link Up Event x1 received Data: x1 xF7 x8 x2 lpfc1:1303:LKe:Link Up Event x1 received Data: x1 xF7 x8 x2

FC - Registered State Change Notification

Fabric Change

vmkernel: 4:00:59:25.152 cpu2)<6>scsi(0): RSCN database changed -0x1,0x400.
vmkernel: 4:00:59:31.574 cpu2)<6>scsi(0): Port database changed.

VMkernel Log: Storage Events

SCSI Reservation conflict

Aug 23 16:42:46 ctc-vmhost-p03 vmkernel: 133:00:48:55.448 cpu5)WARNING: SCSI: 5331: vmhba1:0:8:1 status = 24/0 0x0 0x0 0x0
WARNING: SCSI: 5921: returns 0xbad0023 for vmhba0:0:0

Return codes varies with ESX Server version

Lun/HBA
Non zero value indicates problem

Lun resets

SCSI: 7248: vmhba3:0:3:0 status = 2/0 0x6 0x29 0x0
This translates to: Device Check condition/Host OK the LUN has been reset (bus reset of medium change) device power-on or SCSI reset

Virtual SCSI resets

Aug 8 16:09:31 krraja-dev vmkernel: 48:22:38:34.537 cpu0)SCSI: 2810: Completing reset on handle 62686 (0 outstanding commands)
Aug 8 16:09:45 krraja-dev vmkernel: 48:22:38:48.021 cpu2)SCSI: 771: INQUIRY request with EVDP set

VMkernel Log: Storage Events

Storage Array Mode

May 24 15:51:39 krraja-dev vmkernel: 0:00:00:42.216 cpu4)SCSI: 1464: Device vmhba0:1:0 has not been identified as being attached to an active/passive SAN. It is either attached to an active/active SAN or is a local device

Determines the default Failover policy

Storage Failover

WARNING: LinSCSI: 389: SCSI MODE SENSE command failed with status = I/O error for vmhba1:0:0
WARNING: SCSI: 2812: Manual switchover to path vmhba1:1:0 begins.
WARNING: SCSI: 2360: Did not switchover to vmhba1:1:0. Check Unit Ready Command returned READY instead of NOT READY for standby controller
SCSI: 2816: Changing active path to vmhba1:1:0
WARNING: SCSI: 2841: Manual switchover to vmhba1:1:0 completed successfully.

Virtual Machine Log

- A virtual machine can have more than one process id associated with it.
  - Aug 03 16:45:44: vmx| Log for VMware ESX Server pid=25610 version=2.5.0 build=build-13053 option=Release.2.5.0
  - Aug 03 16:45:44: vmx| Creating thread 'MKS', type 136 from vmware-mks, pid=25611
  - Aug 03 16:47:47: vcpu-0| VMMon_Start: vcpu-0: fd=15 worldID=145
  - Aug 03 16:47:10: vmx| Creating thread 'Floppy', type 16 from self, pid=25761
- Dialogs boxes shown to the user are recorded in the log files.
  - May 02 12:14:55: vmx| [msg.vmxvmdbCb.startInstallTools] Installing the VMware Tools package will greatly enhance graphics and mouse performance in your virtual machine.
  - May 02 12:14:56: vmx| Msg_Question: msg.vmxvmdbCb.startInstallTools reply=1
- Tools version is recorded in the logs
  - Aug 08 11:26:57: vcpu-0| Guest: toolbox: Version: build-13053
- Monitor Panic (Virtual Machine crash) are logged sometimes with the bug number
  - Nov 25 11:29:21: vcpu-0| MONITOR PANIC: BUG F(183):2926 bugNr=27436
- Remote console Connections
  - Oct 05 14:41:16: vmx| VUI: A new gui connected (user = root) (ip = 10.16.12.197).
  - Oct 05 14:41:16: vmx| Accepted new connection at 19 for thread servercontrol (0x8273038)
  - Oct 05 14:41:16: vmx| VUI: A new VMControl client connected.

COS Logs

- /var/log/messages
  - Tracking reboot and configuration changes

    Aug 29 15:26:13 krraja-dev kernel: Kernel command line: auto BOOT_IMAGE=esx ro root=342 mem=192M cpci=0:0,4,6,14,15,16;1:*;2

  - Normal shutdown and startup

    Jun 8 14:48:21 krraja-dev kernel: Kernel logging (proc) stopped.
    Jun 8 14:48:21 krraja-dev kernel: Kernel log daemon terminating.
    Jun 8 14:48:22 krraja-dev syslog: klogd shutdown succeeded
    Jun 8 14:48:22 krraja-dev exiting on signal 15
    Jun 8 14:49:59 krraja-dev syslogd 1.4.1: restart.

  - NMIs passed to the COS are logged: you may not find it if the system crashed before logging

    Jul 25 20:29:00 VA2UVSH06 kernel: NMI received for unknown reason 25.
    Jul 25 20:29:00 VA2UVSH06 kernel: Do you have a strange power saving mode enabled?
    Jul 25 20:29:00 VA2UVSH06 kernel: Please consult hardware error logs

Boot selection

PCI devices assigned to the COS

Normal Shutdown

Logs: More Info…

- Log messages that can be safely ignored are documented in KB articles
  - For ex: KB 1366, 1096, 1292, 1512, 1294, 1296, 1363, 1117
- Certain ALERT messages are important
  - For ex: KB 1081,1300
- Excessive logging need not indicate a problem always. Some drivers tend to be more chatty than others.
- /proc/vmware/log: is similar to dmesg. May contain logs that are not yet flushed to disk
  - Useful in vm-support dumps
- VMkernel logs at the time of PSOD is stored along with the core dump
  - *vmkdump -x zdumpfile* – extracts the VMkernel log file

Proc Nodes

Proc Nodes: IRQ Sharing

- KB 1290: has complete details
- IRQ Sharing between COS and VMkernel can result in performance degradation
- IRQ shared between NIC ports generally doesn't work well
- Disable, USB/Serial port and other unused devices to free up IRQ
- Incorrect MPS table mode setting (KB 1081) can cause IRQ Sharing

 # cat /proc/vmware/interrupts

Vector	PCPU 0	PCPU 1	
0x71:	30	0	COS irq 19 (PCI level), VMK aic7xxx
0x79:	1	52596	<COS irq 17 (PCI level)>, VMK vmnic0
0x81:	66860	0	COS irq 16 (PCI level)

IRQ Shared between COS and VMK!

Used by the COS

Normal. IRQ 17 is not used in COS

Proc Nodes: Guest Timer Interrupts

- /proc/vmware/timers – Timer interrupt state
- http://www.vmware.com/pdf/vmware_timekeeping.pdf
- See KB 892, 1518
- By default, Windows guest demands Timer Interrupts at 100Hz
- Linux 2.6 kernels demand Timer Interrupts at 1000Hz

deadlineTS	periodTS	periodUS	function	data	flags
1686744858945268	747574	500	47415c	0	periodic
1686744870510526	14951486	10000	47b250	9dccb8	one-shot
1686744860003050	14951486	10000	47b250	9e8310	one-shot
1686744867214924	14972418	10014	443ba4	b7	periodic, guest 183
1686744869355698	14951486	10000	47b250	a02c88	one-shot
1686744873607566	14951486	10000	47b250	a0a618	one-shot
1686744859794360	1499634	1003	443ba4	ad	periodic, guest 173
1686744868798312	14972418	10014	443ba4	ab	periodic, guest 171

Guest 183 is demanding Timer interrupts every 10014 Micro seconds, 99 Hz

Guest 173 is demanding timer interrupts at 100000/1003 = 997 Hz

Proc Nodes: Physical NIC

 # watch –d grep –i error /proc/net/PRO_LAN_Adapters/vmnic0.info

Every 2s: grep –i error /proc/net/PRO_LAN_Adapters/vmnic0.info

Mon Oct 3 11:19:58 2005

Rx_Errors	99
Tx_Errors	0
Rx_Length_Errors	0
Rx_Over_Errors	10
Rx_CRC_Errors	77
Rx_Frame_Errors	22
Rx_FIFO_Errors	0
Rx_Missed_Errors	0
Tx_Aborted_Errors	0
Tx_Carrier_Errors	0
Tx_FIFO_Errors	0
Tx_Heartbeat_Errors	0
Tx_Window_Errors	0
Rx_Long_Length_Errors	0
Rx_Short_Length_Errors	0
Rx_Align_Errors	0
Rx_CSum_Offload_Errors	0
PHY_Idle_Errors	0
PHY_Receive_Errors	0

Proc Nodes: Virtual NIC

/proc/vmware/net/vmnicN/macaddr

Virtual Machine 147
Device bond0
PromiscuousAllowed No

World id

Receive queue overflow value

pktsTx	KBTx ….	pktsRx	KBRx…	TxQOvQ	RxQOv ….
Total: 25062	7166	330359	27898	0	0

- For network connectivity issues in a VM check if both pktsTx and pktsRx counters change
- RxQOv: Receive Queue Overflow Value. If this counters increases then the guest is receiving packets at a faster rate than what it can receive
- To sniff traffic on the virtual switch

```
# insmod vmxnet_console devName=vmnicN<.vlanid>
# echo “PromiscuousAllowed yes” > /proc/vmware/net/vmnicN/config
# Ifconfig eth1 up
# tcpdump eth1
```

Proc Nodes: SCSI

/proc/vmware/scsi/vmhbaN/tgt:lun

- contains statistics per lun and per world Id on that lun

```
Vendor: HITACHI   Model: DF600F          Rev: 0000
Type:   Direct-Access                    ANSI SCSI revision: 04
Id: 44 36 30 48 30 31 34 45 30 30 30 31 20 20 20 20 44 46 36 30 30 46
Size:   32770 Mbytes
Queue Depth: 30
…….
Virtual Machine Shares   cmds  reads  KBread  writes  KBwritten   cmdsAbrt   busRst. .. queued
……
Active: 0  Queued: 0
```

Queued should be 0

- Commands queued for a long period indicates the guest is having a Storage I/O bottleneck
- FC-HBA stats can be found at /proc/scsi/lpfc*/ or /proc/scsi/qla*/

Proc Nodes: FC

QLogic PCI to Fibre Channel Host Adapter for QLA2340:
Firmware version: 3.03.08, Driver version 7.04.00
Total number of active commands = 0
Total number of interrupts = 5467
Total number of queued commands = 0

Commands Queued at the FC level

Device queue depth = 0x20

Number of loop resyncs = 0

Number of retries for empty slots = 0

Link State

Host adapter:loop state= <READY>, flags= 0x860813

Commands retried with dropped frame(s) = 0

SCSI LUN Information:

(Id:Lun) * - indicates lun is not registered with the OS.

(0: 0): Total reqs 6, Pending reqs 0, flags 0x0, 1:0:81,

Queued command for the Lun

Proc Nodes: Memory

/proc/vmware/mem

Unreserved machine memory: 520 Mbytes/782 Mbytes
Unreserved swap space: 84 Mbytes/1012 Mbytes
Reclaimable reserved memory: 0 Mbytes
Machine memory free: 672 Mbytes/831 Mbytes
Shared memory (shared/common): 1192 Kbytes/64 Kbytes
Maximum new 1-vcpu virtual machine size: 532 Mbytes
Maximum new 2-vcpu virtual machine size: 524 Mbytes
System heap size: 32768 Kbytes (33554432 bytes)
System heap free: 19229 Kbytes (19691184 bytes)

Memory available for virtual machines

Heap should sufficiently free for stable operation

Don't rely on guest memory usage ! Guest cant tell if the memory pages are from the swap file or reclaimed from another virtual machine

/proc/vmware/sched/mem - per virtual machine memory stats

vm mctl? wait shares min max size/sizetgt memctl/mctltgt swapped/swaptgt

- Size <> sizetgt: Indicates memory pressure. Size > sizetgt : virtual machine is ballooning out. Size < sizetgt: virtual machine is reclaiming memory
- Swappped: Currently swapped to VMkernel page file
- Mctl: Balloon driver availability in the guest – Tools installed ?
- Wait: yes/no - VM waiting for memory !

Proc Nodes: CPU

 # cat /proc/vmware/sched/ncpus

4 logical
2 physical
2 cores
2 sockets

Hyperthreading Enabled

Status: RUN, READY, WAIT, WAITB, ZOMBIE

 # cat /proc/vmware/vm/137/cpu/status

vcpu	vm	type	name	uptime	status	costatus	usedsec	syssec
137	137	V	vmm0:Win2K	357.866	RUN	RUN	265.143	3.105

wait	waitsec	cpu	affinity	htsharing	min	max	shares	emin	extrasec
NONE	51.783	0	0,1	any	0	200	2000	72	124.758

Wait: NONE, IDLE, FS, RQ, RPC, SWPA, SWPS

Effective Minimum % allocation

Performance Issues

Performance Issues: Challenges

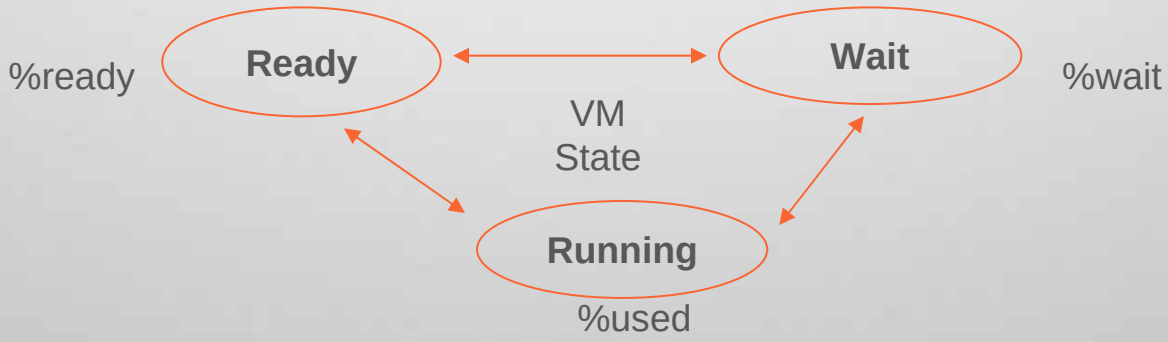
- Symptoms are often misleading or they do not manifest clearly
  - Wrong path of troubleshooting
- Performance depends on a number of factors all of which change constantly
  - Test results may not be consistent
  - Controlled experimentation is very important.
- Traditional performance troubleshooting methodology doesn't always apply to a virtual machine
  - Difficult to generalize rules and metrics.
  - Issues are highly dependent on the workload and the environment.
- Isolating resource bottleneck is the key
  - Translate performance bottleneck into a system metric

Performance: Simple laws

- Law of conservation of performance
  - Virtualization increases utilization not throughput. It improves efficiency not performance.
  - Consolidation, cost, and performance are interrelated
  - Performance tuning is a balancing act
- Law of relativity
  - Metrics are relative and applies only to the context where it is measured
  - Guest metric does not tell us the state of the Virtual machine
- Understand the architecture
  - Identify what resources the application needs and focus on providing/improving those resources
  - *Fully* understand resource allocation policies in ESX Server

Performance Troubleshooting

- %ready shows up when the scheduler is constrained.
  - VCPU:PCPU ratio exceeds 1 and every VCPU wants a PCPU at the same time
  - incorrect sizing/ over-commitment
  - scheduler constraints due to affinity settings, share values
  - %Used + %Ready = CPU utilization desired by the virtual machine
- % wait = idle time or I/O wait time
- Customize esxtop – man esxtop. Save configuration by pressing 'W' (~/.esxtoprc)



Performance Troubleshooting

- Beware of Normalization difference between different tools MUI, vmkusage and VC: http://www.vmware.com/pdf/mui_vmkusage2.pdf
- Use vm-support –s to capture performance snapshot. Use esxtop –R `pwd` to replay from proc node snapshots
- Guest may never report high CPU utilization if the HAL spins instead of halting – KB 1077, 1730
- I/O is impacted by CPU bottlenecks. Keep an eye on pTx/s,pRx/s and r/s, w/s: higher the number, higher the overhead
- High bandwidth usage need not reflect high throughput. I/O errors could cause more bandwidth to be used at lesser throughput

References

- man pages – *highly recommended*
  - mem, cpu, diskbw, net, hyperthreading
- Resource Management for ESX Server I, II
- ESX Server Best Practices for Performance
- ESX Server resources at VMTN: http://www.vmware.com/support/resources/esx_resources.html

Questions?

VMworld2005
virtualizenow

PAC057-C
Troubleshooting VMware System Problems III

Krishna Raj Raja
VMware

VMworld 2005

virtualize now

las vegas • october 18-20, 2005