Troubleshooting & Support

Ong, Kok Leong
Senior Consultant, VMware ASEAN
e klong@vmware.com
m +60-12-4706070
w http://www.vmware.com/

Nick Foo, VSP, VCP, VTSP, VCI
Senior Technical Consultant
Integrated Global Solutions
e foyye@igsb.com.my
m +60 (12) 3837672
w http://www.igsb.com.my/
Session Topic Map

- VMware Support Services
- Approach to troubleshooting Performance
- Approach to troubleshooting Fault
Session Topic Map

VMware Support Services

Approach to troubleshooting Performance

Approach to troubleshooting Fault
Which issue get logged the most?

The image displays a bar chart showing the number of logs for different issues over the months of January to June. The issues listed include:
- ESX 3i Management
- ESX 3i Firmware
- ESX 3i Installation
- ESX Installation
- SAN
- ESX/VM Stability
- Networking
- Storage General
- ESX General
- VM issues

The y-axis represents the number of logs, ranging from 0 to 8000. Each month has a bar for each issue category, indicating the number of logs for that issue in that month.
Severity 1 – Emergency – Server Down

- ESX or VM is completely inaccessible. This can happen for several different reasons including, but not limited to, server failing to boot, server crashing repeatedly, and loss of connectivity to the server.

Severity 2 – Critical - Productivity Effected

- Server is available but is experiencing problems that effect productivity and jeopardize projects.

Severity 3 – Normal – General Problem

- Problem that happens often enough to be recognized as something that needs to be resolved but not one that is effecting productivity.

Severity 4 – Low – Information Request

- Questions about documentation, procedures and general requests for information.
Help Yourself

Raise SR early
➤ Don’t wait until it becomes a major problem

Don’t abuse P1
➤ You may need it in future

Provide full contact
➤ Your mobile + your direct line

Include log
➤ Log is way more than just logging of error messages.
➤ It gives info on the VMware infrastructure
➤ Support has tool to visualize your infra

Provide details
➤ Screenshots
➤ Steps to reproduce, if possible
➤ Non VMware info that are relevant
Help yourself

Active participation at VMTN
- Lots of issues discussed
- Answer questions, not just ask questions

If it complex & urgent, cc your VMware SE
- SE has internal Support Bridge every Thu 11am

Be proactive
- Monitor your SW + HW for warnings & soft errors

Read The ‘Fantastic’ Manual
- Release Notes for known bugs, KB Articles, white paper

If it involves 3rd party
- Raise the SR with the 3rd party too. Keep all parties in-sync

Be good with Network team, Storage team, etc
- It’s virtual Infrastructure, not virtual Server.
Call Flow

Initiating an SR:
- Web Support
- Phone Support

Follow-on Calls:
- Customer Service Reps (CSRs)
- Technical Support Engineers (TSEs)

Technical Support Engineers (TSEs)
Simple process.
A simple SR can be submitted in <5 minutes.
Collecting Logs – User Interface

Log in to the VI Client as an Administrator

Select: File → Export → Export Diagnostics Data (or Administration → Export Diagnostics Data)

Select servers from which to collect the logs including VC Server

Select “Include information ..” Checkbox

Specify location for storing the files.

If file is too big, use www.yousendit.com
vm-support -s or vm-support -S

- Small s add performance data. It’s bundled with regular data collected by vm-support.
- Capital S collect only performance snapshots without the regular data.
- You may specify the collection duration and interval using the following command-line options: vm-support -s -d <duration> -i <interval>
- Default interval period = 2x the amount of time taken to collect a single snapshot. Recommended value is 10 seconds.

Only used during performance issue

- It is important that you use only when your system is experiencing a performance problem.
- If the snapshot collection interval overlaps with a time when the VM is not experiencing a performance problem, the result may skew the data collected and can make analysis difficult.
Subject Matter Expert Support Model

- Specialization by function
  - Install/Configuration
  - Networking
  - Storage
  - Fault
  - OS
  - System Management
- “New hires” start in the Install/Configuration Solutions Team
- Research Team focuses on complex issues
- High volume areas are isolated
- High value services have focused resources
Global Support Coverage

VMware Support Centers
- Palo Alto, California, USA
- Burlington, Canada
- Cork, Ireland
- Bangalore, India
- Broomfield, CO

VMTN Online
Worldwide web community and knowledgebase available
24/7, 365 days a year
Troubleshooting: feature or bug?
Contents

Approach

Concept

- Metrics for CPU
- Metrics for RAM

Tools & Demo

- esxtop
- VI Admin

For future sessions

- Disk Utilisation
- Network Utilisation
Overall Approach

Physical World

OS

App

Virtual World

VI3 Cluster

Host or Resource Pool

Resource Pool or Host

OS in a VM

App in a VM
Virtualization add more

Physical world
- CPU n RAM metrics of the machine

Virtual world
- CPU n RAM metrics of the VM
- CPU n RAM metrics of the Host. May have multiple Resource Pool
- CPU n RAM metrics of the Cluster
Overall Approach

First Thing First

- Look at the **big picture**. The overall health of the cluster.
- Are all host heavily utilised?
  - If vCenter reports heavily utilized CPU/RAM for all hosts in the cluster, then your only choice is pretty much reduce VM or add host.
- If not…
  - Verify that the all systems in the DRS cluster are carrying load when 1 server overloaded. If they aren’t:
    - increase aggression of DRS algorithm
    - check VM reservations against other hosts in the cluster to ensure migrations will happen.
    - check affinity, anti-affinity
    - check CPU pin
    - ensure all hosts are of equal capability (CPU, RAM, Storage, Network)
What do these 2 charts mean?

- Utilization % describes available capacity on hosts (here: CPU usage low, memory usage medium).

- % Entitled resources delivered: best if all 90-100+.

- The chart does not tell which host has what CPU or RAM utilisation.

This DRS only has 1 host. Yours should be >1.
3 main tools

- vCenter client (VI client): per-host stats and per cluster statistics
- esxtop: per-host statistics
- SDK:
  - allows users to collect only the statistics they want
  - Use the VIM API to access statistics relevant to a particular user
  - Can only access statistics that are exported by the VIM API (same API accessible via esxtop/VI client)

All tools use **same** mechanism to retrieve data (special vmkernel calls)
Performance tools provide first-order indication of problems

These tools are useful in different contexts

- Real-time data and troubleshooting: esxtop
- Historical data: vCenter
- Coarse-grained resource/cluster usage: vCenter
- Fine-grained resource usage: esxtop
- Tools within VM: perfmon, sysinternal, top, application-specific tool

Combine information from various tools to get complete picture
Do not use tools (*perfmon*, *top*) within a guess VM

> Result can be misleading.

**Reasons:**

> Their visibility into resources (CPU, RAM, NIC) is based on the virtual configuration, not real physical hardware.
> They are not aware that hardwares are shared.
> Some of their work is performed by VMkernel, which may do it on other CPU than the one assigned.
> The way in which guest OSes account time is different and ineffective in a VM.

**Measuring within a VM is for relative performance, not absolute**

> Comparing components within the VM
> Measuring external impact (e.g. response time to end user)
Native-VM Comparison Pitfalls

Guest reports clock speed of the underlying physical CPU

➤ Resource pool settings may limit the CPU clock cycles

➤ Guest may not get to use the CPU all the time under contention with other VMs

Guest reports total memory allocated by the user

➤ This doesn’t have to correspond to the actual memory currently allocated by the hypervisor
What Is a Resource Pool?

- A logical abstraction for hierarchically managing CPU and memory resources.
- Used on a standalone hosts or VMware DRS-enabled clusters.
- Provides resources for virtual machines and child pools.
Native-VM Comparison Pitfalls

CPU Utilization accounting

- Single threaded application can ping pong between physical CPUs in VMkernel
- CPU utilization reported in task manager is normalized per CPU
- Windows does not account idle loop spinning
- Some work are “delegated” to VMKernel

Available Memory

- Available memory inside the guest may come from swap on the host
CPU: KEY COUNTERS
CPU & VMkernel Scheduler

ESX is designed to run VMs

**Schedulable entity = “world”**

- VMs are composed of worlds
- SC is a world (has agents like vpxa, hostd)
- Helper Worlds. Assist in VMKernel operation

ESX uses proportional-share scheduler to do RM

- Limits
- Shares
- Reservations

Balanced interrupt processing
CPU & VMkernel Scheduler

World states (simplified view):

- ready = ready-to-run but no physical CPU free
- run or used = currently active and running
- wait = blocked on I/O

Multi-CPU VMs => gang scheduling

- Co-run (latency to get vCPUs running)
- Co-stop (time in “stopped” state)

Quiz: the unit is % or millisecond?
ESX: Main screen of esxtop
Explanation of previous slides

CPU Load Average line
- For all physical CPUs, not per core. So this gives a good starting point.
- 3 numbers. Average in 1 minute, 5 minutes and 15 minutes intervals.
- A load average of 1.00 means that total CPUs are fully utilized. Time to reduce load or add CPU.

PCPU(%) line
- Utilization for the CPU by core and in total.
- The comma-delimited data first displayed shows core utilization, followed by "used total" which averages them.
- Rule of thumb: 60% - 80%, so you have room for spike.

LCPU(%) line
- shows the percentage of CPU utilization per logical CPU. The percentages for the logical CPUs belonging to a package add up to 100 percent.
- This line appears only if hyper threading is present and enabled.

CCPU(%) line
- Percentages of total CPU time as reported by Service Console.
  - us = user time in %
  - sy = system time in %
  - id = idle time in %
  - wa = wait time in %
  - cs/sec = context switch per second
- Use of any third party software, such as management agents and backup agents, inside the service console, may result in high CCPU(%) number.

Not applicable in ESXi.
Explanation of previous slides

**Explanation of the “table”**

- **ID** = World ID (if you expand the line of the group) or Group ID (default value, since default display is not expanded)
- **GID** = Group ID, which can be Resource Pool
- **NWLD** = the number of worlds within the specific Resource Pool. An NWLD of 10 means there are 10 lines if you choose to expand it (by pressing the key “e”)

**% of CPU utilisation is based on 1 core.**

- 100% means 1 core.

**A world called “Idle”**

- The %USED entry of the world Idle displays the amount of CPU cycles that remain unused.
- This number can max out at many hundreds of percentages (1600% for a 4 socket, quad-core box) small numbers here represent heavily loaded systems. Assuming 30% buffer, you want to see a number around 480% on a 16-core box.
- Value <100% utilization means <1 physical core remains for additional work.
## Explanation of previous slides

<table>
<thead>
<tr>
<th>Counter</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>%USED</td>
<td>The percentage of CPU that is used by that world or group. A VM that was provided 2 vCPUs, as an example, can max out at 200% CPU utilization.</td>
</tr>
<tr>
<td>%RUN</td>
<td>% of total time scheduled. Exclude hyper-threading and system time, but <strong>include</strong> %OVRLP time.</td>
</tr>
<tr>
<td>%SYS</td>
<td>Part of %USED. % of time spent by VMKernel on behalf of the Resource Pool (or World, if you expand it) to process interrupts and perform other system activities. Not related to kernel time (system) in Linux or UNIX.</td>
</tr>
<tr>
<td>%WAIT</td>
<td><strong>Include</strong> the time the Resource Pool (or World) was idle.</td>
</tr>
<tr>
<td>%IDLE</td>
<td>Deduct this number from %WAIT to get the real waiting time (e.g. waiting for IO event).</td>
</tr>
<tr>
<td>%RDY</td>
<td>% of time the Resource Pool was ready to run, but ESX did not run it due to constraint. Does not include %MLMTD, so it’s a nett %RDY. Should be less than 5% under healthy condition. If ”%RDY - %MLMTD” is &gt;20%, you may experience CPU contention.</td>
</tr>
<tr>
<td>%MLMTD</td>
<td>% of time when ESX deliberately did not run the Resource Pool because it will exceed the limit. A <strong>high number</strong> here means limit was exceeded.</td>
</tr>
<tr>
<td>%CSTP</td>
<td>Time the world spent in ready, co-deschedule state. Roughly speaking, ESX scheduler deliberately puts a VCPU in this state, if this VCPU advances much farther than other VCPUs in the VM. High %CSTP usually means the VM workload does not use VCPUs in a balanced fashion. The VCPU with high %CSTP is used much more often than the others. Do you really need all those VCPUs? Do you pin the guest application to the VCPUs?</td>
</tr>
</tbody>
</table>

*Only for >1 CPU VM*
Ready Time

Only exist in virtual world, not physical.

Guideline

- Ready time above 10% is worth investigation and may be a sign of an over-utilized ESX.

- Sometimes, the %ready goes over 100. That signifies that the VM is starving and does not have access to the CPU. However, it is fine if this value goes above 100 only for a short period of time.

“SMP” and Ready Time

- Many host CPU cores can be idle and yet a VM is accumulating ready time.

- In multi-processor systems, an additional factor affects ready time. VMs that have been scheduled on a particular CPU will be given a preference to run on the same CPU again. This is because of performance advantages of finding data in the CPU cache.

- So the ESX scheduler may choose to let a few cycles on a CPU stay idle, rather than aggressively move a ready VM to another CPU that is idle.

- Our tests indicate that this results in a higher overall system throughput than aggressive relocation of VM from 1 socket to another.
| %OVRLP | Overlapping time. A world still wants CPU but interrupted by another world. High number normally means ESX is experiencing heavy IO |
| %OVRLP | %USED = %RUN + %SYS - %OVRLP |
| %OVRLP | As a result, the overlap value does not incorrectly inflate %USED. |

%RUN continues to accumulate. But %OVRLP kicks in.
esxtop: expand a Group to drill down

- Hit ‘e’ then enter the GID for the VM.
- See picture for example. GID 30, a Win2K VM, was expanded.
- Esx shows counter data for every world in the group.
  - vmmN: a VMM world, 1 for each vCPU. It performs the majority of the work required to execute and virtualize the guest code (OS, application, and hypervisor).
  - vCPU-N: assist the VMM world for each vCPU. Primarily this work revolves around the virtualization of the IO devices.
  - mks: Mouse, keyboard, and screen interrupt servicing.
  - vmware-vmx: assist in maintenance and communications with other worlds and should not represent a material portion of the group utilization.
CPU Chart in VI client

VI client refresh every 20s
Esxtop refresh every 5s
Quiz: impact on value?
Divide VC value by 4
# CPU: Main Counters in VI Client

<table>
<thead>
<tr>
<th>Level</th>
<th>Counter name in API</th>
<th>Description</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CPU.usagemhz.average</td>
<td>The CPU utilization. The maximum possible value here is the frequency of the processors times the number of cores. Example: a VM using 4000 MHz on a system with four 2 GHz processors is using 50% of the CPU ((4000 \div (4 \times 2000) = 0.5))</td>
<td>megaHertz</td>
</tr>
<tr>
<td>1</td>
<td>CPU.usage.average</td>
<td>The CPU utilization. This value is reported with 100% representing all processor cores on the system. As an example, a 2-way VM using 50% of a four-core system is using two cores worth of load.</td>
<td>%</td>
</tr>
<tr>
<td>2</td>
<td>CPU.reservedCapacity.average</td>
<td>CPU Reserved Capacity</td>
<td>megaHertz</td>
</tr>
<tr>
<td>2</td>
<td>CPU.idlesummation</td>
<td>CPU Idle Notice the unit is millisecond, not % Quiz: in real time graph, what does 10000 ms idle mean?</td>
<td>millisecond</td>
</tr>
</tbody>
</table>
CPU: TROUBLESHOOTING TIPS
Performance Issues: CPU

High Wait Time

➢ Workload not CPU-intensive. Look at storage or network
➢ Increasing entitlements will not improve performance

High Used Time

➢ If all PCPUs are near 100%, you maybe over-committing CPU resource.
➢ Check RDY% of the groups in the system to verify CPU over commitment.
➢ If some PCPUs stay near 100%, but others are not, there might be an imbalance issue. Check VM CPU affinity settings.
Performance Issues: CPU

High Ready Time

- Indicates contention
- If CPUs are idle, indicates load-balancing constraints
- Manual affinities
- Hyper-threading sharing constraints
- Hyper-threading Quarantine
  - Default settings allow a VM to share the physical core with all other VM.
  - ESX watches for potentially HT Hostile workloads, and will attempt dynamic quarantine as needed.
  - Non-default settings (Internal or None) can have performance implications especially on low-CPU-count systems
  - Check that HTQ = N in esxtop

Should you use HT or not...?
Performance Issues: CPU

High Co-stop

➤ Co-scheduling overhead
➤ Consider reducing number of virtual CPUs

High CPU utilisation within a VM

➤ As seen by %USED in esxtop
➤ Ensure latest VMware Tools installed. It has paravirtualised driver which reduces CPU load
➤ Consider adding vCPU or RAM (if it decreases disk or network load)
High Ready Time = CPU Overcommit?

2:01:53pm up 4 days 29 min, 87 worlds; CPU load average: 0.16, 0.16, 0.09

D CPU(%): 13.20, 15.55, 10.71, 23.06; used total: 15.63
CCPU(%): 0 us, 0 sy, 99 id, 0 wa; cs/sec: 98

<table>
<thead>
<tr>
<th>ID</th>
<th>GID</th>
<th>NAME</th>
<th>NWLD</th>
<th>%USED</th>
<th>%RUN</th>
<th>%SYS</th>
<th>%WAIT</th>
<th>%RDY</th>
<th>%MLMTD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>idle</td>
<td>4</td>
<td>337.41</td>
<td>336.34</td>
<td>0.00</td>
<td>0.00</td>
<td>61.66</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>system</td>
<td>6</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>599.97</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>helper</td>
<td>22</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>2199.96</td>
<td>0.11</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>drivers</td>
<td>11</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>1099.98</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>console</td>
<td>1</td>
<td>0.92</td>
<td>0.93</td>
<td>0.00</td>
<td>98.27</td>
<td>0.80</td>
<td>0.00</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>vmkapi</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>200.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>vmware-vmkauthd</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>cpuBurn-CLONE</td>
<td>7</td>
<td>50.11</td>
<td>50.03</td>
<td>0.00</td>
<td>495.74</td>
<td>154.20</td>
<td>152.44</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>fakeDB</td>
<td>7</td>
<td>1.43</td>
<td>1.44</td>
<td>0.00</td>
<td>697.67</td>
<td>0.86</td>
<td>0.00</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>Windows 2003 SP</td>
<td>7</td>
<td>5.10</td>
<td>5.12</td>
<td>0.00</td>
<td>693.01</td>
<td>1.87</td>
<td>0.00</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>SQL2005</td>
<td>7</td>
<td>1.63</td>
<td>1.59</td>
<td>0.03</td>
<td>697.16</td>
<td>1.24</td>
<td>0.00</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>memhog-linux-CL</td>
<td>5</td>
<td>1.17</td>
<td>1.16</td>
<td>0.02</td>
<td>498.29</td>
<td>0.55</td>
<td>0.00</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>cpuBurn-CLONE2</td>
<td>7</td>
<td>1.31</td>
<td>1.29</td>
<td>0.03</td>
<td>698.01</td>
<td>0.70</td>
<td>0.00</td>
</tr>
</tbody>
</table>

High Ready Time

High MLMTD: there is a limit on this VM...

⇒High ready time not always because of overcommitment
3 reasons for high ready time

CPU overcommitment

- Possible solution: add more CPUs or VMotion the VM

Workload variability

- A bunch of VMs wake up all at once
- System may be mostly idle (not always overcommitted) but at that point has high ready time

Reservation set on VM

Example

- 4x2GHz host, 2 vCPU VM, limit set to 1GHz (VM can consume 1GHz)
- Without limit, max is 2GHz. With limit, max is 1GHz (50% of 2GHz)
- CPU all busy: %USED: 50%; %MLMTD & %RDY = 150% [total is 200%, or 2 CPUs]
Quiz: what happens here?

```
root@pey-dhcp27:~
10:24:00am up 16:22, 46 worlds; CPU load average: 2.08, 0.78, 0.29
Pcpu(%)  100.00, 100.00 ;  used total: 100.00
Ccpu(%)  1 us, 2 sy, 97 id, 0 wa ;  cs/sec: 196

<table>
<thead>
<tr>
<th>ID</th>
<th>GID</th>
<th>NAME</th>
<th>NMEM</th>
<th>%USED</th>
<th>%SYS</th>
<th>%OVRLP</th>
<th>%RUN</th>
<th>%WAIT</th>
<th>%RDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>idle</td>
<td>2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>system</td>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>500.00</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>console</td>
<td>1</td>
<td>4.78</td>
<td>0.00</td>
<td>0.01</td>
<td>4.79</td>
<td>93.80</td>
<td>1.47</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>helper</td>
<td>13</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>1300.00</td>
<td>0.02</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>drivers</td>
<td>7</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>700.00</td>
<td>0.02</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>vmware-vmkauthd</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
<td>0.00</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>memhog-linux-sm</td>
<td>7</td>
<td>64.68</td>
<td>0.00</td>
<td>0.16</td>
<td>64.84</td>
<td>477.30</td>
<td>58.32</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>kernelcompile-u</td>
<td>5</td>
<td>64.44</td>
<td>0.01</td>
<td>0.11</td>
<td>64.74</td>
<td>365.69</td>
<td>69.91</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>memhog-linux</td>
<td>5</td>
<td>65.62</td>
<td>0.00</td>
<td>0.14</td>
<td>65.76</td>
<td>378.65</td>
<td>55.93</td>
</tr>
</tbody>
</table>
```
Answer

Specification

- The box has 2 cores. No hyper threading.
- At least 3 VMs are running

Situation

- CPU is saturated. Both are 100% used.
- Both can’t cope. Load average touch 2.08 (1 = 100%)
- High Ready Time & High Used Time
- 3 VMs consuming all CPU. Identical priority. So each get 66% of 2 cores (200%)
- idle world gets 0 %USED
Physical pages
- ESX presents the VMs and SC with physical pages
- These pages may be shared by multiple VMs

Machine pages
- Actual pages allocated by ESX from DIMMs
- Multiple physical pages may map to same machine page. This content-based page sharing feature is called Transparent Page Sharing.
esxtop: Memory

Physical Memory in the box

- Vmkernel managed
- SC (ESX)
- PCI Hole

Configuration:
- Configured
- Allocated
- Active
## Overcommit

Memory overcommit average of the past 1, 5 and 15 minutes are displayed. A value of 0.35 means the memory is overcommit by 35%.

Memory overcommit = ratio of **total requested memory** and the **managed memory** minus 1.

- Total requested memory = VM configured memory + user world memory + the reserved overhead memory.
- Managed memory is memory managed by Vmkernel. It excludes things like PCI Hole and SC.

It’s to over commit, so long transparent page sharing & ballooning support it. Use "SWAP" and "MEMCTL" to find whether you are experiencing memory problems.
**Explanation of previous slides**

<table>
<thead>
<tr>
<th>PMEM</th>
<th>Total physical memory breakdown In MB. Total = SC + VMKernel + Others + Free + PCI Hole</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMKMEM</td>
<td>Memory managed by vmkernel in MB.</td>
</tr>
<tr>
<td></td>
<td>• <code>minfree</code> = the minimum memory that ESX would like to keep free.</td>
</tr>
<tr>
<td></td>
<td>• <code>rsvd</code> = currently reserved by resource pool</td>
</tr>
<tr>
<td></td>
<td>• Memory state has 4 possible values: high, soft, hard, low.</td>
</tr>
<tr>
<td></td>
<td>High: Free memory is &gt;6% of (Total – SC)</td>
</tr>
<tr>
<td></td>
<td>Soft: Free memory is &gt;4%</td>
</tr>
<tr>
<td></td>
<td>Hard: Free memory is &gt;2%</td>
</tr>
<tr>
<td></td>
<td>Low: Free memory is &gt;1%</td>
</tr>
<tr>
<td></td>
<td>• High = memory is not under pressure.</td>
</tr>
<tr>
<td>COSMEM</td>
<td>ESX only. ESXi has no COS</td>
</tr>
<tr>
<td></td>
<td>• <code>free</code> — Amount of idle memory.</td>
</tr>
<tr>
<td></td>
<td>• <code>swap_t</code> — Total swap configured.</td>
</tr>
<tr>
<td></td>
<td>• <code>swap_f</code> — Amount of swap free.</td>
</tr>
<tr>
<td></td>
<td>• <code>r/s</code> is — Rate at which memory is swapped in from disk.</td>
</tr>
<tr>
<td></td>
<td>• <code>w/s</code> — Rate at which memory is swapped to disk.</td>
</tr>
</tbody>
</table>
## Explanation of previous slides

<table>
<thead>
<tr>
<th>PSHARE</th>
<th>Page sharing statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;shared&quot; = &quot;common&quot; + &quot;saving&quot;.</td>
</tr>
<tr>
<td></td>
<td>• shared — Amount of physical memory that is being shared.</td>
</tr>
<tr>
<td></td>
<td>• common — Amount of machine memory that is common across worlds.</td>
</tr>
<tr>
<td></td>
<td>• saving — Amount of machine memory that is saved due to page sharing.</td>
</tr>
</tbody>
</table>

| SWAP   | • curr — Current swap usage in Vmkernel. Should be 0, means no swap at ESX level. |
|        | • target — Where the ESX Server system expects the swap usage to be. If target ≠ curr, then ESX will swap to reach target. |
|        | • r/s — Rate at which memory is swapped in by the ESX Server system from disk. Should be 0. |
|        | • w/s — Rate at which memory is swapped to disk by the ESX Server system. Should be 0 |

<table>
<thead>
<tr>
<th>MEMCTL</th>
<th>Memory Control (Balloon) driver data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• curr — Total amount of physical memory reclaimed using the vmmemctl module. If this number changes since last refresh, then ballooning is happening</td>
</tr>
<tr>
<td></td>
<td>• target — Total amount of physical memory the ESX Server host attempts to reclaim using the vmmemctl module.</td>
</tr>
<tr>
<td></td>
<td>• max — Maximum amount of physical memory the ESX Server host can reclaim using the vmmemctl module.</td>
</tr>
<tr>
<td>Counter</td>
<td>Description</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MEMSZ</td>
<td>the amount of configured guest physical memory</td>
</tr>
<tr>
<td>SZTGT</td>
<td>The amount of machine memory to be allocated. Includes the overhead memory for a VM.</td>
</tr>
<tr>
<td>TCHD</td>
<td>The amount of memory (in MB) that has been touched (recently used) by the VM. In this case &quot;recently&quot; means within a minute or two.</td>
</tr>
<tr>
<td>%ACTV</td>
<td>Percentage of active guest physical memory, current value.</td>
</tr>
<tr>
<td>MCTLSZ</td>
<td>Size of the balloon in this VM The amount of memory that the balloon driver is currently holding, for use by other VMs.</td>
</tr>
<tr>
<td>MCTLMAX</td>
<td>Maximum size of the balloon. This much memory can be claimed by balloon and be swapped by Guess OS. Default value is around 65% of configured RAM.</td>
</tr>
<tr>
<td>OVHD</td>
<td>Overhead. The amount of memory used by the VMkernel to maintain and execute the VM.</td>
</tr>
<tr>
<td>SHRD</td>
<td>The amount of the VM's memory that is shared with other VMs.</td>
</tr>
<tr>
<td>SWR/s</td>
<td>The rate at which memory is being swapped in from disk. High swap rates indicate a need for more memory in the cluster.</td>
</tr>
<tr>
<td>SWW/s</td>
<td>The rate at which memory is being swapped out to disk. High swap rates indicate a need for more memory in the cluster.</td>
</tr>
</tbody>
</table>
For many host metrics, charts combine info from VMs

- Memory granted: total granted memory
- Memory ballooned: total ballooned memory
- Memory active: total active memory
- Memory swapped: total swapped memory

For some host metrics, charts are different from VM metrics

- Memory consumed: total consumed + overhead + system
How much RAM does the app (not the OS) actually use?

- Some OS just use all the RAM without much benefit to app.

How much memory are the VMs actually using?

- It is likely that the OS and applications are only using a small percentage of what the VM was assigned. Check the active and touched memory counters for accurate numbers on guest memory usage.

Is memory short in the host?

- Swapping (SWW/s and SWR/s) is a certain sign of this problem.

Is the collection of all VMs active memory exceeds the total available memory?

- Add the TCHD or %ACTV for all VM.
- If so, then either more memory must be added to the host or VMs must be migrated to another DRS cluster.

Are the guests swapping?

- If the VM has been sized with too little memory then the guest OS will swap inside the VM. This will appear to ESX Server as any other disk activity.

Any NUMA issue?

- Check NUMA migration. NMIG reports total migrations since the VM has been powered on. If this number continues to climb then the VM is being migrated from node to node.
- Does the amount of memory located on a remote NUMA node (NRMEM) remain at a non-zero number? This may be a sign that the VM has been sized to exceed the memory of a single NUMA node. If the VM is using more memory than fits on a single node, some of its memory is certain to be located on a remote node.
Memory issue

Host Swapping
➢ Bad. Avoid this (see value of SWCUR)
➢ Could bt too many VMs or high memory over commit

Guest swapping
➢ Maybe ok (e.g. ballooning). Investigate more.
➢ Memory size for VM is too small?
➢ User processes stealing memory?
➢ Use vmstat inside the guest to detect.

NUMA issues
➢ Memory is not “next to” the CPU.
➢ See the values of NRMEM (Node Remote Memory) & NLMEM (Local)
➢ Seen in larger box, especially 8 sockets
Ballooning & performance

Which of the 2 VMs experience more swapping?

<table>
<thead>
<tr>
<th>NAME</th>
<th>MEMSZ</th>
<th>SZTGT</th>
<th>MCTRL</th>
<th>MCTRLSZ</th>
<th>MCTRLGT</th>
<th>MCTRLMAX</th>
<th>SWCUR</th>
<th>SWTGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows 2003 SP</td>
<td>1024.00</td>
<td>385.53</td>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
<td>665.60</td>
<td>119.82</td>
<td>0.00</td>
</tr>
<tr>
<td>SQL2005</td>
<td>2048.00</td>
<td>456.73</td>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
<td>1331.20</td>
<td>215.91</td>
<td>0.00</td>
</tr>
<tr>
<td>vc server</td>
<td>1024.00</td>
<td>284.19</td>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
<td>665.60</td>
<td>78.65</td>
<td>0.00</td>
</tr>
<tr>
<td>fakeDB</td>
<td>2048.00</td>
<td>483.75</td>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
<td>1331.20</td>
<td>203.79</td>
<td>0.00</td>
</tr>
<tr>
<td>memhog-linux-sm</td>
<td>1024.00</td>
<td>376.21</td>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
<td>665.60</td>
<td>628.47</td>
<td>620.38</td>
</tr>
<tr>
<td>memhog-linux-CL</td>
<td>1024.00</td>
<td>347.39</td>
<td>N</td>
<td>652.21</td>
<td>652.21</td>
<td>652.21</td>
<td>55.10</td>
<td>57.07</td>
</tr>
</tbody>
</table>

Memory Hog VMs
MCTRL: N - Balloon driver not active, tools probably not installed
Balloonning active
Swapped in the past but not actively swapping now
More swapping since balloon driver is not active
Answer: the VM with no Memory Control driver

Configuration
➢ The box has only 4 GB of physical RAM
➢ 6 VMs are running. Total configured virtual RAM is 8 GB

What happen
➢ 2 VMs asking for 100% of their configured VM.
   ▶ Shown as last bottom 2 VM.
   ▶ Name begin with “memhog” for memory hogging.
➢ Not enough physical RAM. So swapping is unavoidable.

How Ballooning helps
➢ VM 1 has no balloon driver. ESX has no visibility on the RAM inside the VM, and do swapping at VMkernel level. VM does not know that ESX swaps; it thinks all its RAM are on DIMM. Result? Worse performance.
➢ VM 2 has balloon driver. The Guess OS moves inactive pages to disk, doing its own swapping. This reduces pressure to ESX. As a result, ESX does less swapping.
Type of Faults

Possible faults/bugs

- Hardware
- ESX: SC, VMM, VMKernel
- VM
- HA & DRS
- Network
- Storage
- vCenter

For future sessions

- Troubleshooting HA n DRS
- Troubleshooting Network
- Troubleshooting Storage
- Troubleshooting vMotion
ESX fault

Typically classified as a fault with the VMkernel or SC

- Vmkernel faults manifests in a Purple Screen crash (PSOD)
- SC crashes result of Linux kernel Oops or panic
- Entire host may hang
- Spontaneous reboots of host
A PSOD (Purple Screen Crash) is when the VMkernel enters a condition where it cannot or should not proceed and causes a crash.

Common purple screen crash types are:

- CPU Exceptions
- SC Oops/Panic
- Machine Check Exception (MCE)

A purple screen fault is usually caused by either a hardware problem or a genuine bug in the system.
VMware ESX Server [Releasebuild-64607]

Exception type 13 in world 1082:vmn0:backup- @ 0x7cbde4
frame=0x3aeba04 ip=0x7cbde4 cr2=0xfd8c9000 cr3=0x366b2000 cr4=0x760
es=0x30204041 ds=0x72204041 fs=0x72750000 gs=0x676e4041
eax=0x58eb825e ebx=0x1 ecx=0x3 edx=0x2e56ad
ebp=0x3aebad8 esp=0x3f8c2e00 edi=0xa834700 err=0 eflags=0x11212
0:1278/vmm0:svn 1:1111/mks:weble *2:1082/vmm0:back 3:109/vmm0:pro1
0x3aebad0:[0x7cbde4]SCSICompletePathCommand+0x1f stack: 0x3f8c2e00, 0x86e, 0xa
0x3aebbf8:[0x7cb201]SCSIStartPathCommand+0x85c stack: 0xfd62, 0x0, 0x8db5706f
0x3aebc64:[0x7cc5e7]SCSIIssueAsyncPathCommand+0x14a stack: 0x0, 0xa834700, 0x538
2880
0x3aebd10:[0x7d3560]SCSI legacy MPC issueCommand+0x7 stack: 0x0, 0xa834700, 0x20
0x3aebd50:[0x7d459e]SCSI_LegacyMDeviceStart+0x241 stack: 0x3f008100, 0x3f008100
0x834700
0x3aebd60:[0x7bd377]SCSIStartDevice+0x6e stack: 0x3f008100, 0x8c6f24, 0x43a
0x3aebd68:[0x7bd3c6]SCSIAsyncDeviceCommand+0x2bd stack: 0x3f8d9a00, 0xa834700, 0x0
0x3aeebe0c:[0x7b595d]SCSI_IssueAsyncDeviceToken+0x1cc stack: 0x3f8d9a00, 0x3ea561
0, 0x8
0x3aeebe4c:[0x79653d]USCSI RawDiskCommand+0x1f8 stack: 0x6629480, 0x985c000, 0x3aebe0
0x3aeebf2c:[0x790168]USCSIHandleCommand+0x32f stack: 0x3aeebf7c, 0x6, 0xa05e000
0x3aeebf50:[0x791790]USCSIExecCommandInt+0x103 stack: 0x3aeebf7c, 0x6, 0xa05e000
0x3aeebf8c:[0x7919e9]USCSI_ExecuteCommand+0x12c stack: 0x53, 0x8000000, 0x5954
0x3aeebfe0:[0x621e15]VMKCall+0x12c stack: 0x53, 0x5954, 0x8000000
VMK uptime: 75:16:05:22.201 TSC: 13043151362457774
Starting coredump to disk using slot 1 of 1... 9876666543210 Disk dump successful,
Debugger is listening on serial port ...
This section shows the type of crash that has occurred

- Shows which VMkernelworld (or VM) caused the crash
- An exception type 13 (Page Fault) has occurred in world 1082
- World 1082 is the VM whose name starts with “backup-”
This section shows what was running on each CPU

- The CPU with an asterisk before it is the CPU which the crash occurred on.
- Format: <CPUnumber>: <World ID>/<name of world>
- The fault occurred on CPU 2 world 1082 running a VM whose name starts with “back”
SC Oops/Panic

PSOD caused by a SC Linux kernel Oops or panic
Separate stack trace from the Linux kernel displayed after the regular PSOD message

Be sure that the following advanced options are set correctly:

➤ Misc.PsodOnCosPanic
➤ Misc.CosCoreFile
Beyond the Purple Screen

How to find out more?

- Look for vmkernel-zdump-* file in /root
  - Use vmkdump–l <zdumpfile> to extract the log from the dump
- Review the vmkernel-log.1 file for more information.
  - Go to the end of file. Look at purple screen output
  - Read log entries before purple screen for clues to problem
  - For SC Oops/panic, the Vmkernel log will contain:
    - First half is from the vmkernel
    - Second half is from the SC Linux kernel
Beyond the Purple Screen

Example vmkernel-log.1 output:

75:07:20:58.992 CPU1:1237)DevFS: 2208: Unable to find device: vmdisk-000002-delta.vmdk
75:07:21:03.274 CPU3:1237)DevFS: ...
75:16:05:22.199 CPU2:1082)WARNING: ScsiPath: 2158: Can't split CMD 0a, returning Illegal Request to initiator 0x6629700

ESX Server [Releasebuild-64607]
Exception type 13 in world 1082:vmm0:backup- @ 0x7cbde4
frame=0x3aeba04 ip=0x7cbde4 cr2=0xdf8c9000 cr3=0x3c6b2000 cr4=0x760
es=0x30204041 ds=0x72204041 fs=0x72750000 gs=0x676e4041
eax=0x58eb825e ebx=0x1 ecx=0x3 edx=0x2e56ad
ebp=0x3aebad8 esi=0x3f8c2e00 edi=0xa834700 err=0 eflags=0x11212
0:1278/vmm0:svn 1:1111/mks:weble *2:1082/vmm0:backup 3:1109/vmm0:prol
@BlueScreen: Exception type 13 in world 1082:vmm0:backup- @ 0x7cbde4
0x3aebad8:[0x7cbde4]SCSICompletePathCommand+0x1f stack: 0x3f8c2e00, 0x86e, 0xa
0x3aebbf8:[0x7cb201]SCSIStartPathCommands+0x85c stack: 0xfd62, 0x0, 0x8db5706f
0x3aebc64:[0x7cc5e7]SCSIAsyncPathCommand+0x14a stack: 0x0, 0xa834700, 0x5382880
0x3aebd18:[0x7d3568]SCSIAsyncDeviceStart+0x7f stack: 0x0, 0xa834700, 0x20
0x3aebd50:[0x7d459e]SCSI_LegacyMPDStart+0x241 stack: 0x3f008100, 0x3f008100, 0xa834700
0x3aebd68:[0x7b3e77]SCSIAsyncDevice+0x6e stack: 0x3f008100, 0xa8c6f24, 0x43a
Beyond the Purple Screen

Other considerations

➢ For storage related issues, consider setting up serial-line logging

➢ If purple screen shows that a memory core dump has not been successfully written to disk, try setting up a vmkcore partition on a remote storage like a FC LUN (for local storage issues)

➢ For SC Oops/panic

  ▪ Look for any offending processes that may have caused the issue in the vmkernel log output

  ▪ If third-party software has been installed, consider disabling to see effects
Thank you

Ong, Kok Leong
Senior Consultant, VMware ASEAN
e klong@vmware.com
m +60-12-4706070
w http://www.vmware.com/

Nick Foo, VSP, VCP, VTSP, VCI
Senior Technical Consultant
Integrated Global Solutions
e fooye@igsb.com.my
m +60 (12) 3837672
w http://www.igsb.com.my/