Improving Guest Operating System Accounting for Descheduled Virtual Machines in ESX Server 3.x Systems

The experimental VMware Descheduled Time Accounting component, VMDesched, is an optional new component of VMware Tools. VMDesched is available starting with ESX Server 3.0.

Note: Currently, VMDesched is available only for uniprocessor Windows and Linux guest operating systems.

When installed, the VMDesched component provides two key benefits:

- Improved accuracy for guest operating system CPU time accounting when physical CPU resources are overcommitted. Time that a virtual machine spends descheduled is now represented explicitly within the guest by a VMDesched process or thread. This enables performance monitoring tools and workload management software to account for descheduled time accurately, preventing other guest processes and threads from being charged incorrectly for this time.
- Improved guest operating system timekeeping with respect to real time. Running VMDesched helps a virtual machine that has been descheduled catch up to real time quickly and safely after it resumes execution.

This document explains how to install and monitor VMDesched on Linux and Windows guest operating systems. It also describes timer interrupt virtualization issues resolved by VMDesched and how VMDesched works.

Installing VMDesched on Linux Guest Operating Systems

Linux Guest Operating System Requirements: The current release supports all uniprocessor Linux 2.4.x and 2.6.x kernels.

To install VMDesched on Linux Guest Operating Systems

1. Install VMware Tools as specified in *Basic System Administration*.

   Note: You must install VMware Tools before you can activate VMDesched.

2. Run the `vmware-config-tools.pl` script with the `--experimental` argument to install experimental features:

   ```
   # vmware-config-tools.pl --experimental
   ```

3. When prompted to install the VMware Descheduled Time Accounting daemon, choose `yes`.

   If a pre-built module exists for your kernel, that module is installed. If a pre-built module does not exist, you are prompted to build one as you were for the other VMware Tools modules during the full VMware Tools installation.

Once VMDesched is installed, you can start and stop it automatically with the rest of the VMware Tools using the VMware Tools `/etc/init.d/vmware-tools` initialization script.

To monitor descheduled time within the Linux guest operating system, inspect the `vmware-vmdesched` process.

Installing VMDesched on Windows Guest Operating Systems

Windows Guest Operating System Requirements: The current release supports all uniprocessor versions of Windows 2000, Windows XP, and Windows Server 2003.

To install VMDesched on Windows Guest Operating Systems

1. Start the VMware Tools installation as specified in *Basic System Administration*.
2. Select Custom as the Setup Type option.

   The VMDesched component is not installed by default.
3. In the Custom Setup program features menu, expand the VMware Device Drivers option menu.
4. Click the drop down menu next to the Descheduled Time Accounting item and choose This feature will be installed on local hard drive.
5. Complete the installation, and reboot the guest operating system.

Once VMDesched is installed, you can start VMDesched manually by issuing the following command from the Windows command prompt:

```
net start vmdesched
```

You can also start VMDesched automatically, as described in the following section.

To monitor descheduled time within the Windows guest operating system, inspect the `vmdesched.exe` process, using Task Manager or another monitoring tool.

To start VMDesched automatically

1. From the Windows Start menu, choose Settings > Control Panel > Administrative Tools > Services.
2. Right-click the VMware Descheduled Time Accounting service entry and choose Properties to open the Properties page.
3. Set the Startup Type option to Automatic, and click Apply.

Timer Interrupt Virtualization Issues Resolved by VMDesched

Most operating systems, including Windows and Linux, depend on periodic timer interrupts for two important activities:

- The operating system advances its notion of real time in response to each timer interrupt. Elapsed real time since the operating system was booted is often maintained as a simple count of timer ticks.
- The operating system performs statistical process accounting by charging the process that was running while the timer interrupt occurred for consuming a timer tick's worth of CPU time.

Note: The generic term *process* refers to an entity that can be scheduled within the guest operating system. Different guests may use terms including *process*, *thread*, or *task*.

When an operating system is running within a virtual machine as a guest operating system, both of these activities can be affected by the virtualization of timer interrupts. In particular, when virtual machines are running on a single physical machine, processor resources are time-multiplexed, with the result that some virtual machines can be descheduled for relatively long periods of time. When the virtual machine resumes execution after being descheduled, it might have accumulated a backlog of timer interrupts corresponding to the elapsed real time during which it was descheduled. For example, with a timer period of 10 milliseconds, if a virtual machine is descheduled for 50 milliseconds, it will be "behind" by 5 timer interrupts when it resumes execution.

By default, the VMware virtualization layer helps a guest that is behind "catch up" to real time by delivering its backlog of timer interrupts at a faster rate. This might distort statistical process accounting in the guest operating system, because descheduled time is attributed to those processes that are running when the catch-up interrupts are delivered. For more details, see the white paper *Timekeeping in VMware Virtual Machines,* available at *http://www.vmware.com/pdf/vmware_timekeeping.pdf*.

How VMDesched Works

The VMDesched VMware Tools component coordinates with the VMware virtualization layer to determine when there is a backlog of timer interrupts. The virtualization layer arranges to deliver backlogged timer interrupts to the virtual machine only while the guest operating system is executing the VMDesched process. The backlog is cleared as quickly as allowed by the guest operating system.

Because all backlogged timer interrupts are delivered in the context of the VMDesched process, the guest operating system effectively charges VMDesched for consuming all of the time when the virtual machine was descheduled. This allows performance monitoring tools—such as `top` in Linux and Task Manager in Windows—to report virtual machine descheduled time without distorting the CPU time charged to actual guest processes. This also enables third-party performance tools and workload management software to detect virtual machine descheduled time and to incorporate this information when generating reports or alerts related to utilization.

VMware, Inc. 3145 Porter Drive Palo Alto, CA 94304 www.vmware.com