Setting up VMWare ESXi with NVIDIA GRID for MAK software

MAK software products function in multiple layers. Virtualization technology enables sharing of high performance resources, hardware cost reduction, distributed management, and security, among other benefits.

One configuration includes the VMWare ESXi bare metal hypervisor managing virtual machines on local hardware, or in a cloud configuration. NVIDIA GRID allows multiple virtual machines to have direct access to a single physical GPU using normal drivers, and VMWare Horizon Client provides remote desktops on networked hardware.

VMWare provides vSphere to manage virtual machines on an ESXi host. If you wish to run multiple VMWare ESXi hosts, as with a cloud configuration, VMWare vCenter Server provides a single interface to conveniently manage multiple vSphere instances at extra cost.

NVIDIA GRID technology requires the CPU to fully support Intel Virtualization Technology, and Intel VT for Directed I/O (VT-d). Ensure that both features are enabled in the host machine's BIOS. Links to support information are in the **Resources** section.

Document sections:

This document assumes the following computer hardware and software.

Hardware Specifications

AMAX ServMax 2U Intel Xeon Server consisting of:

- Intel Server System R2312WF0ZSSPP, 2U 12 x 3.5" Bay
- BIOS: AMAX Performance Optimized
- CPU: 2x - Intel® Xeon® Platinum 8160 Processor - 24core 2.1Ghz
- RAM: 4x - 32GB DDR4 2666MHz x4 DR RDIMM CT32G
- GPU: 2x - Nvidia Tesla P40 24GB GDDR5

Software Specifications

GPU Software:

- NVidia GRID version 390.72
- NVidia Driver version 391.81

Virtual Machine Software:

- VMWare vSphere/ESXi 6.7 with VMWare Tools
- VMWare vCenter Server
- VMWare Horizon Client (optional)

Guest Operating System for Virtual Machines:

- Microsoft Windows 10 (64-bit)

Setting up the VMWare ESXi Hypervisor Host

Before you begin:

Ensure the host machine BIOS confirms that both Intel Virtualization Technology, and Intel VT for Directed I/O (VT-d), are enabled.

Download VMWare ESXi, vSphere Server, and VMWare Tools.

Download the NVIDIA Virtual GPU Manager that supports your VMWare version.

Steps:

1. Install VMWare ESXi on the host machine.
   1.1. Create bootable media (USB or DVD) for VMWare ESXi.
   1.2. Set the host machine to boot from your chosen media.
   1.3. Restart the host machine, and run the installer when it appears.
   1.4. Follow instructions to complete installation.
2. Set up hostname, datastore, and networking, according to VMware documentation.
3. Prepare Guest Operating System installation media to create virtual machines.
   3.1. Upload ISO installation images to a datastore on the host.
      Note: You can instead create physical ISO installation disks that must be placed in the host machine for each use.
4. Install and configure NVIDIA Virtual GPU Manager for VMWare vSphere on the ESXi host according to VMWare documentation.
5. Access the ESXi host to confirm that the NVIDIA Virtual GPU Manager installed successfully, obtain ESXi host GPU details, and disable ECC memory support.

5.1. Enable Secure Shell (SSH) from the vSphere Web Client.

5.1.1. Select the host machine, then click **Manage,** and keep **Settings** selected.

5.1.2. Click **Security Profile**.

5.1.3. In the **Services** section, click **Edit**.

5.1.4. Select **SSH,** and click **Start** to begin the SSH Service.

Note: This setting does not persist after you reboot the host. To persist SSH Service after you reboot, select **Start and stop with host** and then reboot.

5.1.5. Click **OK** to accept the setting.

5.1.6. Use the **Actions** menu to confirm that SSH is enabled, as shown in the following image. When SSH is enabled, **Disable Secure Shell (SSH)** is visible under the **Services** menu item.

[Image 1. vSphere Actions menu showing Secure Shell (SSH) is enabled.]

5.2. Open a connection to the ESXi host with an SSH client such as PuTTY.

5.2.1. Enter the IP address of the ESXi host machine as **hostname.**

5.3. Confirm successful installation of the NVIDIA Virtual GPU Manager.

5.3.1. Enter the command: `vmkload_mod -l | grep nvidia` and a successful output resembles: `nvidia  310 13816` as shown in the following image.

[Image 2. PuTTY SSH client showing NVIDIA Virtual GPU Manager installed]

5.4. Display status and detailed information of NVIDIA GPU software and drivers associated with the current ESXi host. One of these values, the `Bus-Id` can be used to adjust a specific GPU.

5.4.1. Enter the command: `nvidia-smi` and the output displays the host system time followed by NVIDIA details as shown in the following image.

[Image 3. PuTTY SSH client showing NVIDIA GPU detailed information]

5.5. You must ensure that ECC memory support is disabled for any NVIDIA GPU that drives virtual GPUs, or they will fail to start. This instruction shows how to toggle the `Ecc Mode` for all GPUs, or for a single GPU with use of the `Bus-Id`.

5.5.1. Enter the command exactly: `nvidia-smi -q | grep Ecc -A 2` and the output displays only the `Ecc Mode` section of the full `nvidia-smi -q` output, including two lines indicating `Ecc Mode` status, shown in the following image.

[Image 4. PuTTY SSH client showing NVIDIA ECC memory support Disabled]

5.5.2. Disable ECC memory support for all GPUs when you enter the command: `nvidia-smi -e 0`

5.5.3. Disable ECC memory support for a single GPU when you add `-i Bus-Id` before `-e 0` in the previous command. For example, Image 3. shows `GPU 0` with a `Bus-Id` of `00000000:18:00.0`. To disable ECC memory support for only `GPU 0`, enter the command: `nvidia-smi -i 00000000:18:00.0 -e 0`

5.5.4. Reboot the ESXi host.

5.5.5. When the ESXi host is rebooted, ensure that SSH is enabled, connect with an SSH client and verify that ECC memory support is disabled when you enter the command exactly: `nvidia-smi -q | grep Ecc -A 2`

Result: The VMWare ESXi host should now be ready to add virtual machines.

Setting up a Virtual Machine (VM) on the VMWare ESXi Host

Before you begin:

Consider the resource needs of your environment. VT MAK provides hardware recommendations with other information to make your setup easier and more successful. Our goal is get your application working and we are always here to help.

Steps:

1. Log in to vSphere Server for the ESXi host machine and follow VMWare documentation to create a VM with the suggested configuration.

Guest OS:	Win 10 (64-bit) VBS not enabled
Compatibility:	VM version 14
VMWare Tools:	Yes
CPU:	8 vCPUs, 1 socket, Reservation 16760 MHz
Memory:	32768 MB
Storage:	1 disk
HD 1:	150 GB, Thick provisioned, lazily zeroed
PCI device 0:	NVIDIA GRID vGPU
GPU Profile:	p40-8q

Note: To set up the correct NVIDIA Virtual GPU Type, please review the VT MAK website: http://www.mak.com/support/hardware-recommendations

The following screenshots show suggested virtual machine configuration settings.

[Image 5. VMWare vSphere VM CPU configuration screenshot]

[Image 6. VMWare vSphere VM Memory configuration screenshot]

[Image 7. VMWare vSphere VM HDD configuration screenshot]

[Image 8. VMWare vSphere VM Networking configuration screenshot]

[Image 9. VMWare vSphere VM External Device configuration screenshot]

[Image 10. VMWare vSphere VM Video Card configuration screenshot]

Setting up NVIDIA licensing for NGRID technology

The NVIDIA GRID Software and NVIDIA License Manager are downloaded from NVIDIA's Licensing Portal. Once installed, the GRID License Server provides an MAC address, used to generate licenses with NVIDIA's Licensing Portal. The resulting license `.bin` file is uploaded to the GRID License Server, where the Guest OS retrieves a license during boot to activate Tesla GPU features, and returns it during shutdown.

Before you begin:

Determine where to run your GRID License Server. If run on the same machine with NVIDIA GRID-based VMs, the host system CPU and memory resources are utilized. Plan your hardware accordingly. A standalone license server has its own resources.

Steps:

1. Access the NVIDIA Licensing Portal website to obtain software.
    1.1. Log in, or Register with your Product Activation Key (PAK).
    1.2. Go to the NVIDIA Software Licensing Center and click your GRID version link.
    1.3. Download the GRID Software and NVIDIA License Manager for your version.
2. Create a License Server interface.
    2.1. Unzip and install the GRID License Server at your determined location.
    2.2. Copy and save the MAC address as provided by the GRID License Server.
    2.3. Return to the NVIDIA Software Licensing Center website.
    2.4. Select **Register License Server** to display the Create Server web page.
    2.5. Enter the recorded MAC address and server details before you click **create**.
    2.6. When complete, select the created server to display the View Server page.

2.7. **IMPORTANT**: The license file is blank by default. Select **Map Add-Ons** to associate the required number of licenses to your server before you download the license file. Refer to VMWare documentation for any updated details.

2.8. **IMPORTANT**: The `.bin` license file is only valid for 24 hours from when it is created and must be uploaded to the License Server before the time expires. Select **Download License File** and save the resulting `.bin` license file locally.

2.9. Go to http://<FQDN of the license server>:8080/licserver to display the License Server Configuration page, and then select **Configuration** from the left menu.

2.10. Select **Choose File** to open a file browser, and upload the `.bin` license file.

3. Connect the Guest OS to a running GRID License Server and check out a license.

3.1. Launch a virtual machine and log in to it.

3.2. Right click on the desktop and open the NVIDIA Control Panel. (See image 11.)

3.3. Find the **Manage Licenses** section.

3.3.1. Enter the GRID License Server name, and port 7070, where appropriate.

3.3.2. Ensure the VM checks out a valid license from the GRID License Server.

3.3.2.1. Watch for an error message indicating that a license is not found.

3.3.2.2. When no valid GRID license is found, MAK products (VR-Engage, VR-Forces, and VR-Vantage) are restricted to 3 FPS.

[Image 11. NVIDIA Control Panel showing the Manage License section]

Result: The VM with running MAK products use the NVIDIA GRID-based GPU at normal speed without error messages.

Resources

VT MAK Software

https://www.mak.com/support/product-user-guides

VMWare

ESXi

https://www.vmware.com/products/esxi-and-esx.html

vSphere

https://www.vmware.com/products/vsphere.html

Horizon

https://www.vmware.com/products/horizon.html

NVIDIA

GRID Software Documentation

https://docs.nvidia.com/grid/index.html

GRID Quick Start Guide

http://images.nvidia.com/content/pdf/grid/guides/quickstartguide.pdf

GRID VGPU installation

https://docs.nvidia.com/grid/latest/grid-GPU-user-guide/index.html