S9299 NVIDIA VGPU ON RED HAT LINUX HYPERVISOR (RHV)

Shailesh Deshmukh
Senior Solution Architect, Nvidia

Konstantin Cvetanov
Senior Solution Architect, Nvidia

Eric Kana
Senior Solution Architect, Nvidia

Sal Lopez
Senior Solution Architect, RedHat

AGENDA

- *What We Will Discuss*
- *Benefits of Virtualization*
- *Why RHV and vGPU?*
- *Target Audience and Prerequisites*
- *Design & Implementation*
- *Demonstration*
- *Summary*

WHAT WE WILL DISCUSS

A practical approach to design and implement NVIDIA Virtual GPU on Red Hat Hypervisor in an enterprise IT environment

…and why you would want to do it!

BENEFITS OF VIRTUAL DESKTOP INFRASTRUCTURE

- Enable flexible workflow scenarios
- Utilize centralized, shared, and protected storage
- Enable intellectual property protection
- Provide flexibility in configuration
- Enable user/workforce mobility
- Widely supported GPU acceleration

What you planned the system to do.

NVIDIA vGPU SOLUTION

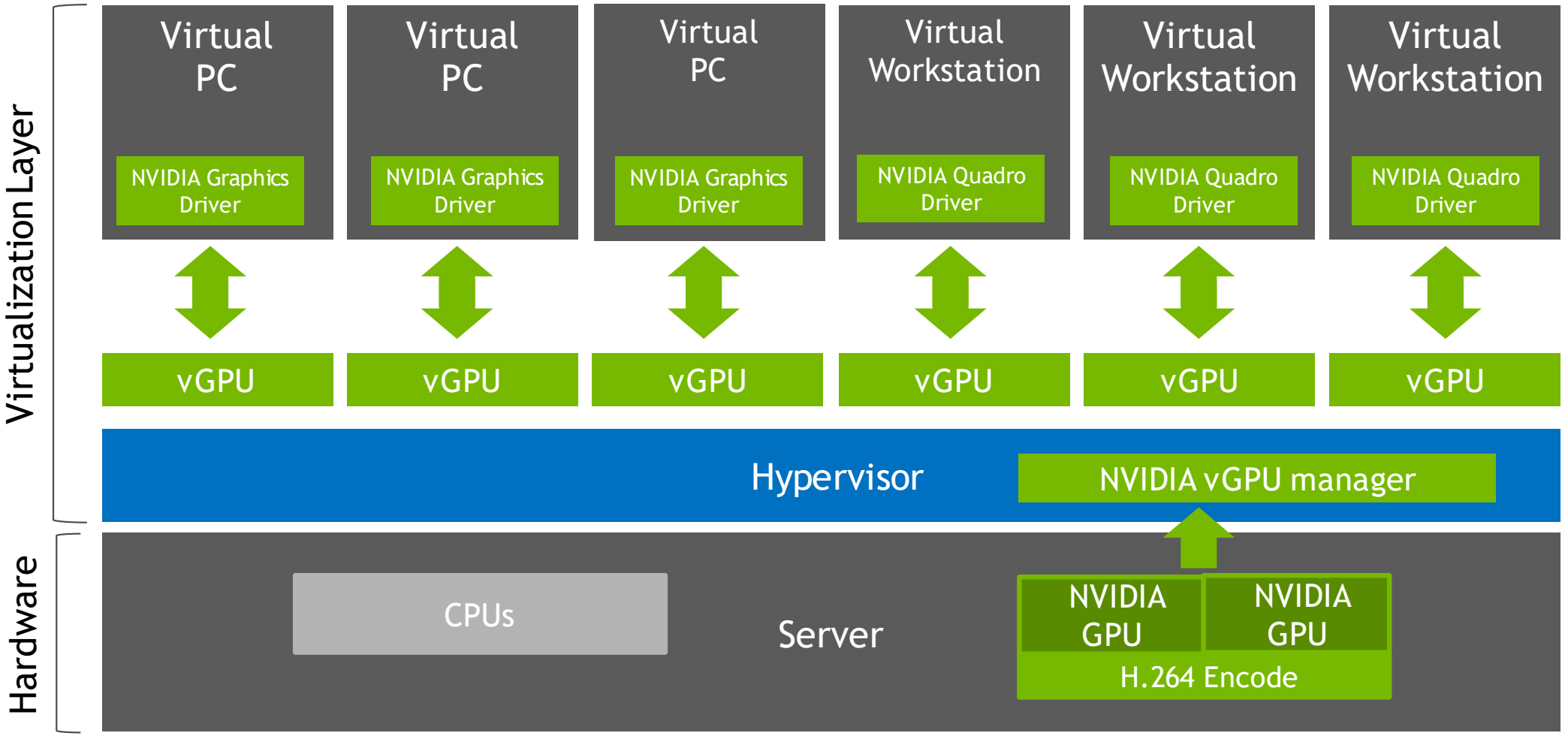
Industry Standard Graphics Virtualization Platform

Support, Updates & Maintenance

NVIDIA Virtual GPU Software

Nvidia Data Center GPUs

HOW DOES NVIDIA vGPU WORK?



WHY RED HAT VIRTUALIZATION?

Maximizes physical infrastructure utilization while reducing costs

- RHEL is the leading Linux distro for enterprises
- Supports both graphics and compute workloads in both Linux and Windows
- Utilize one GPU across several virtual technical workstations using mediated device support
- Support for multi-vGPU implementations

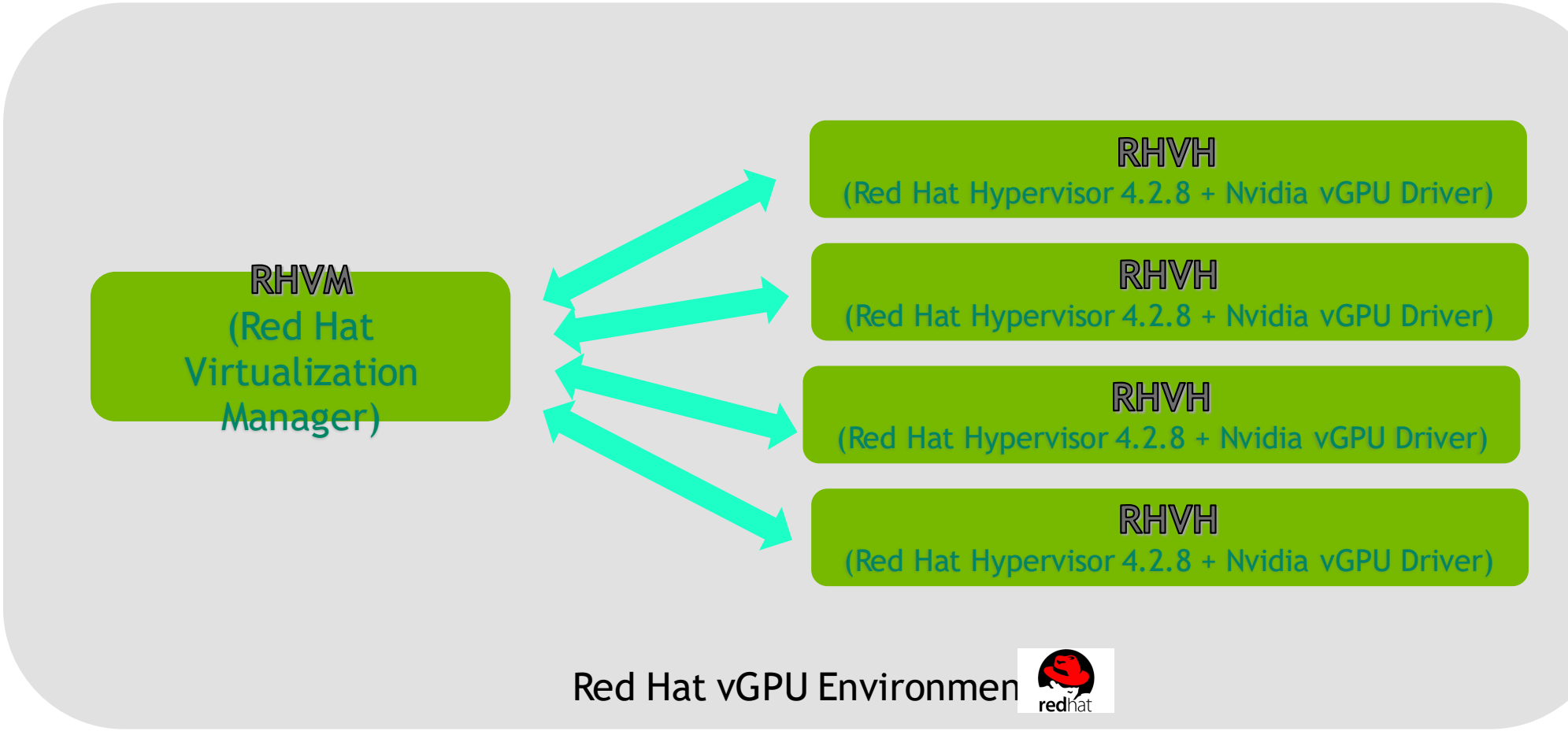
Automation and Seamless Deployments

- Centralized management in a data center deployment
- Customer can re-use many RHEL 7 security practices for their RHV infrastructure
- Tighter integration with the Red Hat portfolio via consistent API's vs competitors portfolio and disparate proprietary plugins

Wide Customer Adoption

- Oil & Gas, Energy, Animation, Manufacturing and Gaming

REFERENCE ARCHITECTURE



BEFORE YOU START

- Target Audience
  - ❑ Linux System Administrators
  - ❑ Infrastructure Architects
- Hardware Requirements
  - ❑ Supported Server OEM Platform
  - ❑ Supported NVIDIA Data Center GPU (no K1/K2)
- Software Requirements
  - ❑ RHV 4.2.8 or later ISO
  - ❑ RHEL 7.5 or later ISO
  - ❑ NVIDIA vGPU 7.0 or later
  - ❑ NVIDIA License Server

VIRTUAL GPU MANAGER

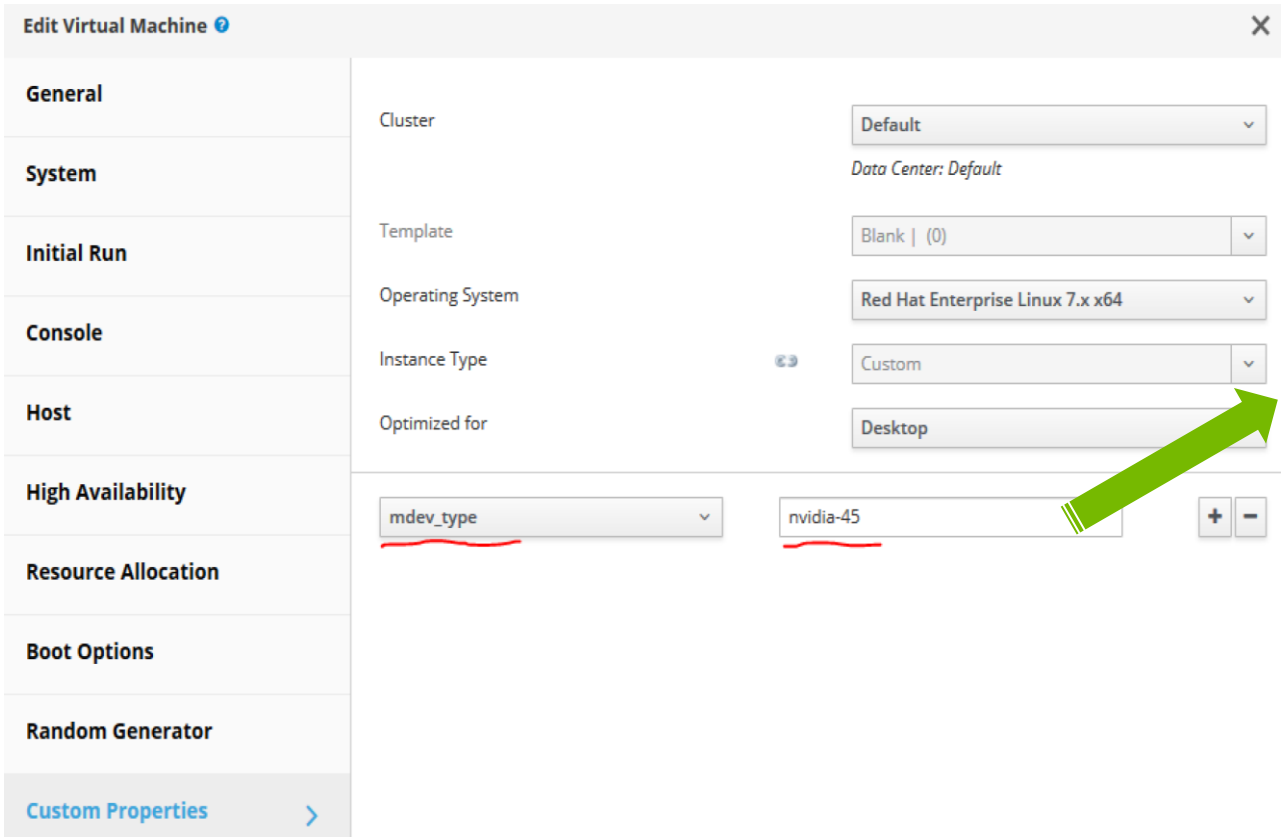
Download, deploy, verify vGPU Manager:

```
[root@rhvh tmp]# rpm -ivh NVIDIA-vGPU-rhel-7.6-410.91.x86_64.rpm
Preparing...                          ################################# [100%]
Updating / installing...
   1:NVIDIA-vGPU-rhel-1:7.6-410.91    ################################# [100%]
[root@rhvh tmp]#
```

```
[root@rhvh ~]# nvidia-smi
Tue Mar 12 13:31:23 2019
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 410.68       Driver Version: 410.68       CUDA Version: N/A      |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla M10           On   | 00000000:06:00.0 Off |                  N/A |
| N/A   41C    P8    10W /  53W |     11MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla M10           On   | 00000000:07:00.0 Off |                  N/A |
| N/A   41C    P8    10W /  53W |     11MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   2  Tesla M10           On   | 00000000:08:00.0 Off |                  N/A |
| N/A   40C    P8    10W /  53W |     11MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
|   3  Tesla M10           On   | 00000000:09:00.0 Off |                  N/A |
| N/A   40C    P8    10W /  53W |     11MiB /  8191MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

VIRTUAL MACHINE CONFIGURATION (1)

Add vGPU profile as mdev device:



```
"mdev": {
    "nvidia-155": {
        "name": "GRID M10-2B",
        "available_instances": "4"
    },
    "nvidia-36": {
        "name": "GRID M10-0Q",
        "available_instances": "16"
    },
    "nvidia-37": {
        "name": "GRID M10-1A",
        "available_instances": "8"
    },
    "nvidia-35": {
        "name": "GRID M10-0B",
        "available_instances": "16"
    },
    "nvidia-38": {
        "name": "GRID M10-1B",
        "available_instances": "8"
    },
    "nvidia-39": {
        "name": "GRID M10-1Q",
        "available_instances": "8"
    },
    "nvidia-43": {
        "name": "GRID M10-4Q",
        "available_instances": "2"
    },
    "nvidia-42": {
        "name": "GRID M10-4A",
        "available_instances": "2"
    },
    "nvidia-41": {
        "name": "GRID M10-2Q",
        "available_instances": "4"
    },
    "nvidia-40": {
        "name": "GRID M10-2A",
        "available_instances": "4"
    },
    "nvidia-45": {
        "name": "GRID M10-8Q",
        "available_instances": "1"
    },
    "nvidia-44": {
        "name": "GRID M10-8A",
        "available_instances": "1"
    },
    "nvidia-208": {
        "name": "GRID M10-2B4",
        "available_instances": "4"
    }
```

NVIDIA VGPU PROFILES EXPLAINED

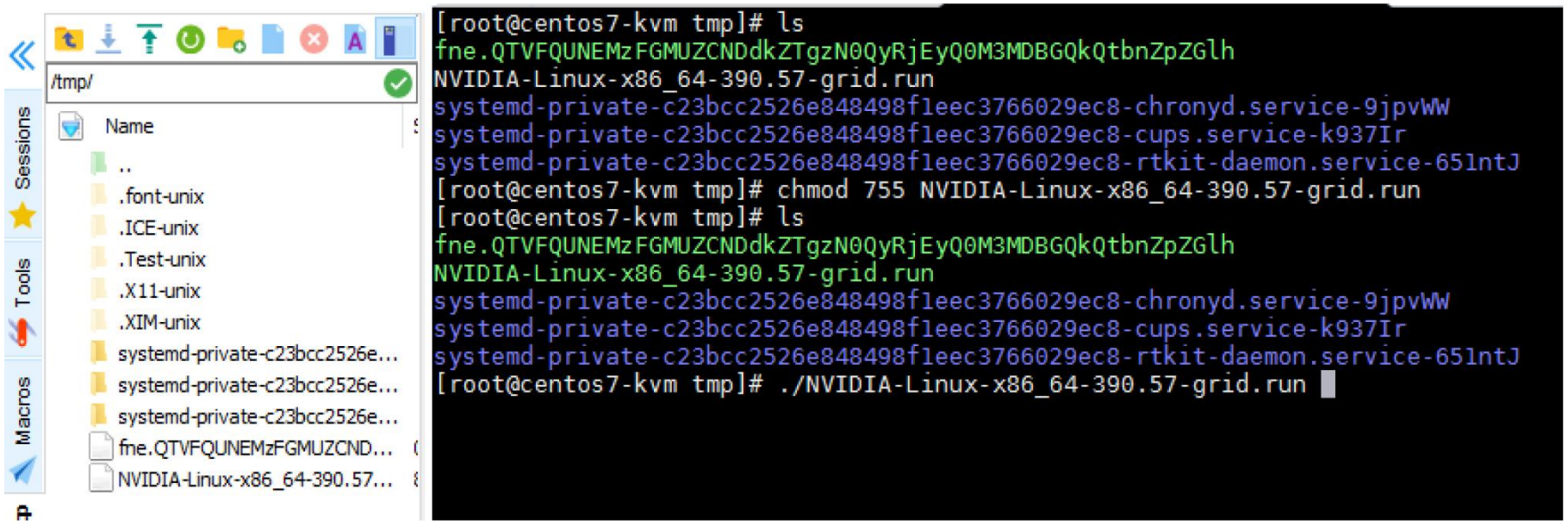
Q - Profiles	Quadro Features Enabled
B - Profiles	Basic Display Driver
A - Profiles	Application Profile

1Q is 1 Gig of Frame Buffer
2B is 2 Gig of Frame Buffer
8A is 8 Gig of Frame Buffer

Xq/b/a → “X” is the amount of Frame Buffer on the Nvidia GPU card

VIRTUAL MACHINE CONFIGURATION (2)

Inside the Linux guest VM install NVIDIA Virtual GPU driver:



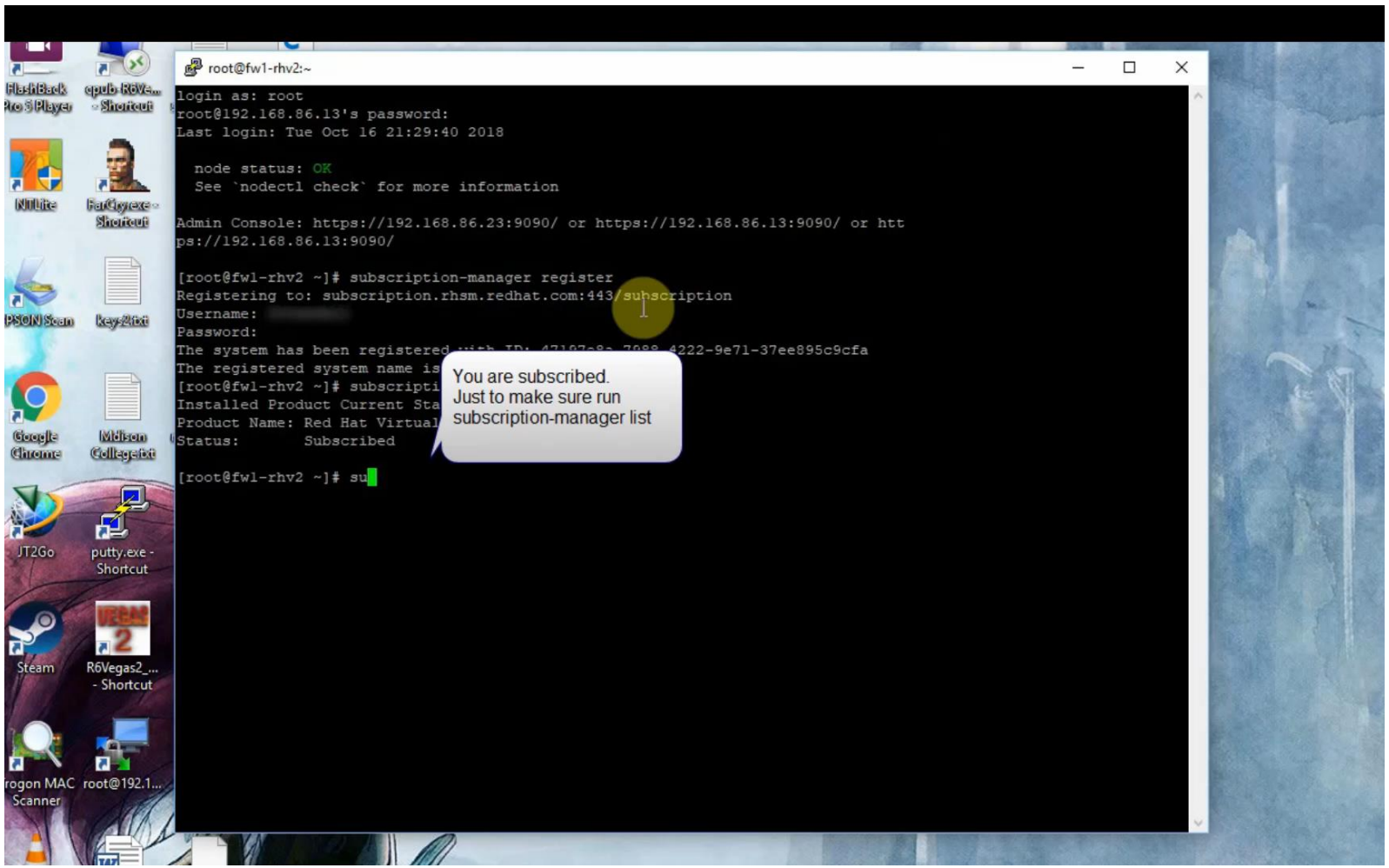
LICENSE SERVER CONFIGURATION

Create a License Server VM and edit the grid.conf file:

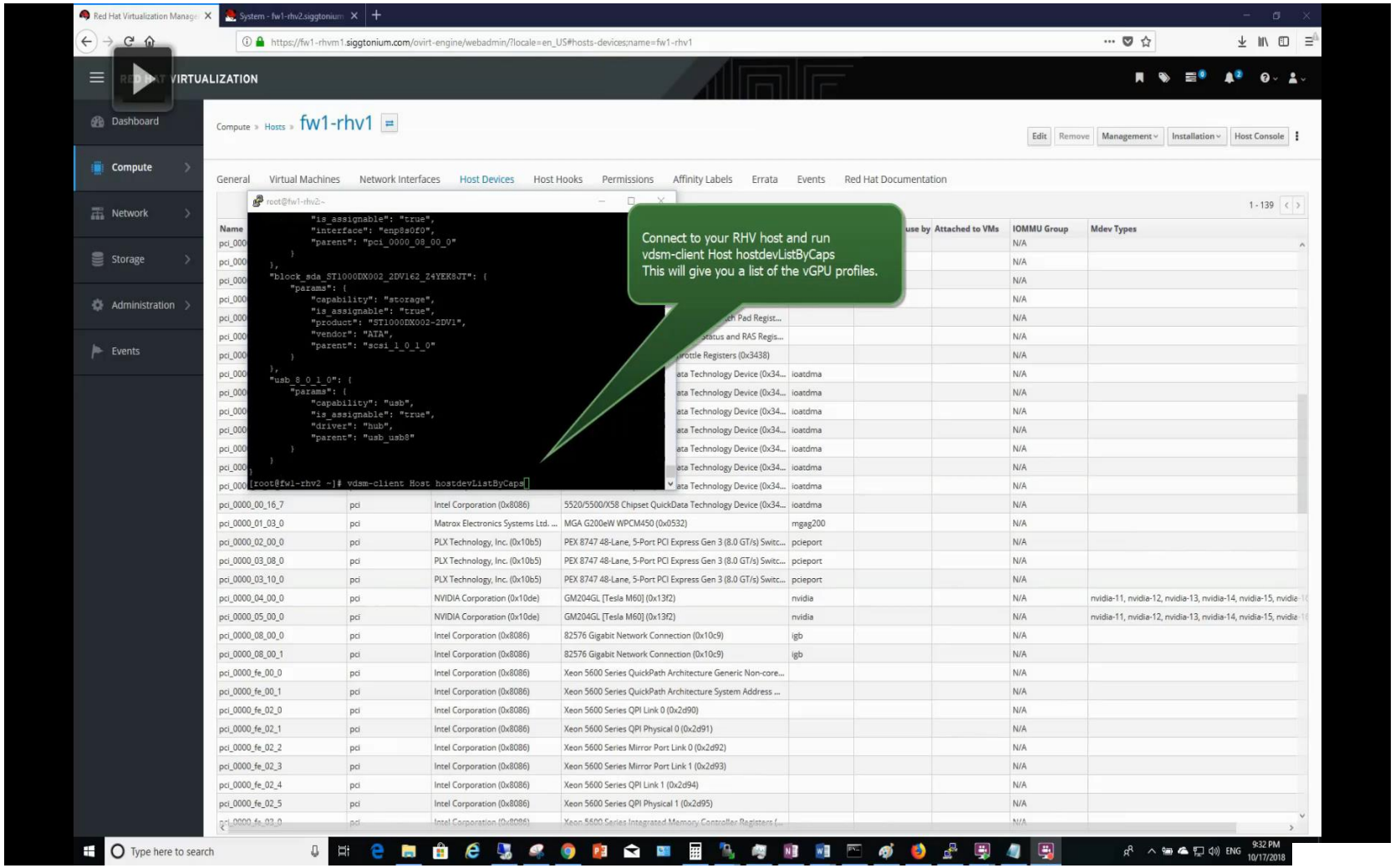
```
# For details on the file format, please refer to the nvidia-gridd(1)
# man page.

# Description: Set License Server Address
# Data type: string
# Format:  "<address>"
ServerAddress=10.31.230.12

# Description: Set License Server port number
# Data type: integer
# Format:  <port>, default is 7070
ServerPort=7070

# Description: Set Backup License Server Address
# Data type: string
# Format:  "<address>"
#BackupServerAddress=

# Description: Set Backup License Server port number
# Data type: integer
# Format:  <port>, default is 7070
#BackupServerPort=

# Description: Set Feature to be enabled
# Data type: integer
# Possible values:
#    0 => for unlicensed state
#    1 => for GRID vGPU
#    2 => for Quadro Virtual Datacenter Workstation
FeatureType=1

# Description: Parameter to enable or disable Grid Licensing tab in nvidia-settings
# Data type: boolean
# Possible values: TRUE or FALSE, default is FALSE
#EnableUI=TRUE

# Description: Set license borrow period in minutes
# Data type: integer
# Possible values: 10 to 10080 mins(7 days), default is 1440 mins(1 day)
#LicenseInterval=1440

# Description: Set license linger period in minutes
# Data type: integer
# Possible values: 0 to 10080 mins(7 days), default is 0 mins
#LingerInterval=10
[root@centos7-kvm tmp]#
```

VIRTUAL GPU MANAGER

Download, deploy, verify vGPU Manager:



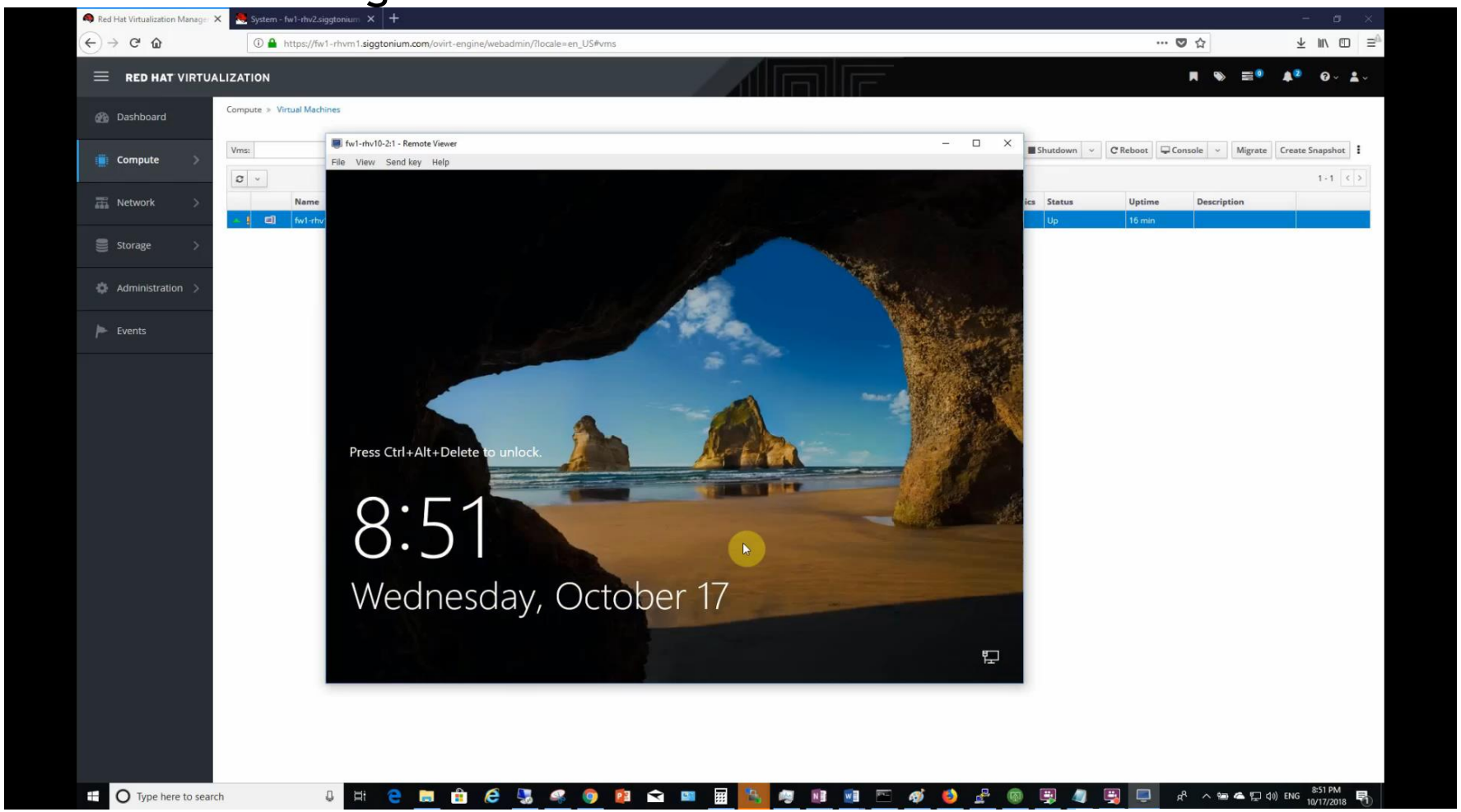
VIRTUAL MACHINE CONFIGURATION




```
"mdev": {
    "nvidia-155": {
        "name": "GRID M10-2B",
        "available_instances": "4"
    },
    "nvidia-36": {
        "name": "GRID M10-0Q",
        "available_instances": "16"
    },
    "nvidia-37": {
        "name": "GRID M10-1A",
        "available_instances": "8"
    },
    "nvidia-35": {
        "name": "GRID M10-0B",
        "available_instances": "16"
    },
    "nvidia-38": {
        "name": "GRID M10-1B",
        "available_instances": "8"
    },
    "nvidia-39": {
        "name": "GRID M10-1Q",
        "available_instances": "8"
    },
    "nvidia-43": {
        "name": "GRID M10-4Q",
        "available_instances": "2"
    },
    "nvidia-42": {
        "name": "GRID M10-4A",
        "available_instances": "2"
    },
    "nvidia-41": {
        "name": "GRID M10-2Q",
        "available_instances": "4"
    },
    "nvidia-40": {
        "name": "GRID M10-2A",
        "available_instances": "4"
    },
    "nvidia-45": {
        "name": "GRID M10-8Q",
        "available_instances": "1"
    },
    "nvidia-44": {
        "name": "GRID M10-8A",
        "available_instances": "1"
    },
    "nvidia-208": {
        "name": "GRID M10-2B4",
        "available_instances": "4"
    }
```

VIRTUAL MACHINE CONFIGURATION (2)

Inside the guest VM install NVIDIA Virtual GPU driver:




Thank you
NVIDIA