Utilizing NVIDIA GPUs in Virtualized Deployments with Red Hat’s Portfolio

Martin Tessun - Senior Technical Product Manager, Red Hat
Michael Shen - Senior Product Manager, NVIDIA
Shawn Hall - HPC Systems Engineer, Numerical Algorithms Group

May 2019
NVIDIA and Red Hat partnership
Where Red Hat Partners with NVIDIA

- GPU accelerated workloads in the enterprise
  - AI/ML and HPC
- Deploy and manage NGC containers
  - On-prem or public cloud
- Managing virtualized resources in the data center
  - vGPU for technical workstation
- Fast deployment of GPU resources with Red Hat
  - Easy to use driver framework
Red Hat/NVIDIA Technology Partnership Timeline

Nvidia GTC 2017: Red Hat vGPU Roadmap Update

-May'17
- STAC-A2 Benchmark (Nvidia/HPe/RHEL-STAC Conf NYC, RH & Nvidia Blogs)

-Nov'17
- SC2017 RH/Nvidia (booth demos, talks)

-Mar’18
- Nvidia GTC 2018 & Kubernetes WG mtg - RH vGPU & Kubernetes sessions, RH sponsorship

-Apr’18
- LSF & MM Summit: Nouveau Driver demo

-May’18
- RH Summit - AI booth & OpenShift Partner Theatre & RH AI/ML Strategy sessions

-Jun’18
- vGPU/RHV - Joint Webinar - Oil & Gas Use Case

-Oct’18
- NVIDIA GTC DC; RHEL & OpenShift Certification on DGX-1

-Dec’18
- OpenShift Commons/KubeCon; Deep Learning on OpenShift w/GPUs

-Mar’19
- NVIDIA GTC 2019; RHEL & OpenShift Certification on DGX-2 / T4 GPU Server Configs, RH Sponsorship

NVIDIA GTC 2017: Red Hat vGPU Roadmap Update

-May'17
- STAC-A2 Benchmark (Nvidia/HPe/RHEL-STAC Conf NYC, RH & Nvidia Blogs)

-Nov'17
- SC2017 RH/Nvidia (booth demos, talks)

-Mar’18
- Nvidia GTC 2018 & Kubernetes WG mtg - RH vGPU & Kubernetes sessions, RH sponsorship

-Apr’18
- LSF & MM Summit: Nouveau Driver demo

-May’18
- RH Summit - AI booth & OpenShift Partner Theatre & RH AI/ML Strategy sessions

-Jun’18
- vGPU/RHV - Joint Webinar - Oil & Gas Use Case

-Oct’18
- NVIDIA GTC DC; RHEL & OpenShift Certification on DGX-1

-Dec’18
- OpenShift Commons/KubeCon; Deep Learning on OpenShift w/GPUs

-Mar’19
- NVIDIA GTC 2019; RHEL & OpenShift Certification on DGX-2 / T4 GPU Server Configs, RH Sponsorship

NVIDIA GTC 2017: Red Hat vGPU Roadmap Update

-May'17
- STAC-A2 Benchmark (Nvidia/HPe/RHEL-STAC Conf NYC, RH & Nvidia Blogs)

-Nov'17
- SC2017 RH/Nvidia (booth demos, talks)

-Mar’18
- Nvidia GTC 2018 & Kubernetes WG mtg - RH vGPU & Kubernetes sessions, RH sponsorship

-Apr’18
- LSF & MM Summit: Nouveau Driver demo

-May’18
- RH Summit - AI booth & OpenShift Partner Theatre & RH AI/ML Strategy sessions

-Jun’18
- vGPU/RHV - Joint Webinar - Oil & Gas Use Case

-Oct’18
- NVIDIA GTC DC; RHEL & OpenShift Certification on DGX-1

-Dec’18
- OpenShift Commons/KubeCon; Deep Learning on OpenShift w/GPUs

-Mar’19
- NVIDIA GTC 2019; RHEL & OpenShift Certification on DGX-2 / T4 GPU Server Configs, RH Sponsorship
Red Hat + NVIDIA: What’s New?

- Red Hat Enterprise Linux Certification on DGX-1 & DGX-2 systems
  - Support for Kubernetes-based, OpenShift Container Platform
  - NVIDIA GPU Cloud (NGC) containers to run on RHEL and OpenShift
- Red Hat’s OpenShift provides advanced ways of managing hardware to best leverage GPUs in container environments
- NVIDIA developed precompiled driver packages to simplify GPU deployments on Red Hat products
- NVIDIA’s latest T4 GPUs are available on Red Hat Enterprise Linux
  - T4 Server with RHEL support from most major OEM server vendors
  - T4 servers are “NGC-Ready” to run GPU containers
Red Hat + NVIDIA: Open Source Collaboration

Open Source Projects

• Heterogeneous Memory Management (HMM)
  • Memory management between device and CPU

• Nouveau Driver
  • Graphics device driver for NVIDIA GPU

• Mediated Devices (mdev)
  • Enabling vGPU through the Linux kernel framework

• Kubernetes Device Plugins
  • Fast and direct access to GPU hardware
  • Run GPU enabled containers in Kubernetes cluster
Customer Usecase
Background

- Customer has a single High Performance Computing center in Houston (AKA “the HPC”)
- Researchers globally use the HPC
- Impractical to replicate compute, storage, and networking resources at every major site
- What is easier to move across the network – terabytes of compressed seismic data or compressed pixels?
  - We use Expedat from Data Expedition to move data quickly
- Answer: compressed pixels
Questions from management

- Most researchers already have a nice workstation at their desk, so why don’t we just shove all their workstations in the server room close to the data?
  - Cost inefficient
  - Space inefficient
  - Hard to administer
- Can we just use servers and virtualize them? What do we do with the workstations?
Red Hat Virtualization and vGPU
WHAT IS RED HAT VIRTUALIZATION

Centralized Management of virtualized compute, network, and storage resources using the Open Source KVM Hypervisor

Automated workload management, scalability, and security features for virtualized applications

Engineered to optimize current IT and integrate with future technologies using a RESTful API.
## Virtual Technical Workstation

<table>
<thead>
<tr>
<th>Reduce costs for Technical Workstations</th>
<th>Maximizes physical infrastructure utilization</th>
<th>Fast Deployment and Self Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>● No dedicated hardware per technical workstation is needed</td>
<td>● Supports both Linux and Windows workloads</td>
<td>● Resource Management to buy new hardware in time</td>
</tr>
<tr>
<td>● Centralized Management and Deployment in a Datacenter.</td>
<td>● Utilize one GPU across several virtual technical workstations using mediated device support</td>
<td>● Fast deployment process for new virtual technical workstations including Self Service</td>
</tr>
</tbody>
</table>
VIRTUAL GRAPHICS PROCESSING UNIT

vGPU powered technical workstation support for AI, big data, rich graphics

- NVIDIA (GRID and Quadro vDWS)—maintainer of mediated device framework (mdev)
  - Support multiple vGPU devices per VM (GRID 7.0 and RHV 4.2.7 needed; restrictions to vGPU types apply)
- Intel (GVT-G)—driver development and reviewer for mdev
- Support for Linux and Windows

- Oil and gas
- Energy
- Sciences and education
- Manufacturing and engineering
- Animation
- Gaming

PERFORMANCE SENSITIVE
DEV AND TEST ENVIRONMENTS
HYBRID AND MULTIHYPervisor
TECH WORKSTATIONS
SERVER CONSOLIDATION

#redhat #rhsummit
Use Cases for Multiple GPU with RHV

- HPC, Image Processing & Rendering, Machine Learning, Artificial Intelligence,
  - Oil & Gas: Seismic image rendering
  - Media & Entertainment: Creative design and video editing
  - Automotive: Automotive design, Self driving car
  - Health care: Medical image processing and rendering
  - Research & Education: High performance modeling

- Accelerated performance
- Improved productivity
- Flexibility
- Better TCO
Customer Usecase
Cost comparison

Dell Precision 7820 Workstation
- 1 x Intel Xeon Gold 6140 18C @ 2.3 GHz
- Nvidia Quadro P4000 8 GB
  - Subtract $598 for P400 2 GB
- 96 GB RAM 2666 MHz DDR4
- 2 x 3.5” 1 TB 7200 RPM HDD
- $9,875 (Dell website)
- $945 (remote graphics software)
- Grand total: $10,820

Dell R740 Rack Server
- 2 x Intel Xeon Gold 6140 18C @ 2.3 GHz
- Nvidia Tesla P40 24 GB
- 768 GB RAM 2666 MHz DDR4
- 8 x 3.5” 1 TB 7200 RPM HDD
  - Just use NAS w/ dedup in reality
- $51,936 (Dell website)
- $16,380 (8 VMs - software licensing)
- $43,453 (24 VMs – software licensing)
- Sliced into 8 VMs = $8,540 per user
  - Oversubscribed 2:1 = $4,270 per user
- Sliced into 24 VMs = $3,975 per user
  - Oversubscribed 2:1 = $1,987 per user
Oversubscription benefits of virtualization

- Oversubscribe CPU cores
  - E.g. Give each VM 18 cores on a 36 core host
- Oversubscribe memory*
  - Memory ballooning add/removes memory from VMs based on pressure
- Oversubscribe GPU
  - vGPU memory is fixed, but vGPU can use all host GPU cores
  - Requires vGPU scheduling policy change**
- Oversubscribe network
  - 40/100 GbE host network shared by VMs

* In theory – has not been successfully tested w/ vGPU
Components used at customer

- Server with Nvidia Tesla series GPU
  - Tesla GPU required for GPU virtualization
- Nvidia vGPU capable hypervisor
- Nvidia vGPU host and guest drivers
- Mechdyne TGX remote graphics software
  - Captures video of remote system and sends to local system
- Leostream Connection Broker software
  - Acts as traffic officer directing users to available systems
Nvidia vGPU

- Gives greater flexibility for choosing slices of GPU allocated to VMs
- Fewer GPU cards with multiple GPU chips per card makes vGPU a necessity
NVIDIA Virtual GPU Technology
How Does NVIDIA vGPU Work?

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual PC
- NVIDIA Graphics Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Virtual Workstation
- NVIDIA Quadro Driver
- vGPU

Hypervisor
- NVIDIA vGPU manager

Server
- NVIDIA GPU
- NVIDIA GPU
- H.265 Encode
NVIDIA vGPU Solution
Industry Standard GPU Virtualization Platform

Support, Updates & Maintenance

NVIDIA Virtual GPU Software

Nvidia Data Center GPUs

#redhat #rhsummit
NVIDIA T4 FOR VIRTUALIZATION
Universal GPU for Everyone

Virtual Quadro Workstation for the Professional Designer & Data Scientist:
• Up to 2X graphics performance versus M60
• 5 Giga Rays per second for real-time, interactive rendering
• NGC support; run deep learning inferencing workloads 25x faster than CPU on a virtual machine

Virtual PCs for the Knowledge Worker:
• Support for VP9 decode and H.265 encode and decode for improved CPU offload
NEW Quadro virtual Data Center Workstation
(Release April 2019):

- RTX 6000 and RTX 8000
- Up to 1.4X graphics performance than P40
- RTX Server with Virtual Workstation
Multi-vGPU
Available on RHV

- Multiple VM’s Sharing a GPU with NVIDIA Quadro vDWS
- A Single VM Accessing Multiple NVIDIA GPUs with NVIDIA Quadro vDWS (7.0 or later)
“The flexibility of the new multi-GPU feature available with NVIDIA Quadro vDWS opens up powerful new rendering workflows to SOLIDWORKS Visualize users. The near linear performance scaling means they can iterate on their designs at lightning speed on professional virtual workstations, allowing our customers to arrive at their best design in the shortest amount of time.” - Brian Hillner, SOLIDWORKS Product Portfolio Manager
New host support
- RHEL 8 / RHV 4.3
- XenServer 8.0

New guest support
- Windows 10 RS5
- Windows Server 2019
- RHEL 8

New NVIDIA GPU support
- RTX 6000 / 8000

New CSP Marketplace support
- GCP
- Azure
- Alibaba Cloud
Customer Use Case
Mechdyne TGX

- Offloads computation through Nvidia Capture SDK to GPU to minimize latency
- High color accuracy
- Multi user collaboration
- Data is SSL encrypted
- Compatible with Windows, Mac, and Linux
- Performance improvements come “automatically” with GPU improvements
Leostream

- Very flexible, vendor agnostic connection broker
- Directs users to available systems based on policies
- Uses 3 components
  - Leostream Connection Broker – routes users to a desktop, manages TGX connection settings, defines policies around desktop usage
  - Leostream Connect – runs on user’s computer, establishes and configures TGX session based on Connection Broker response
    - Windows, Mac, Linux clients
  - Leostream Agent – runs on remote computer and helps the Connection Broker to manage connections
Architecture overview
Architecture detail

Leostream Connect

Which system?
Use this one.

Leostream Connection Broker

TGX Receiver

System and configuration info
Please start / resume a session for me

NVDEC

Client

Keyboard and mouse inputs

Capture GPU frame buffer

X / Gnome session

startx

TGX Sender

H.264 video w/ AAC audio

VM w/ vGPU

NVENC

CUDA

Client

System and configuration info

Leostream Connect

Which system?
Use this one.

Leostream Connection Broker

TGX Receiver

System and configuration info
Please start / resume a session for me

NVDEC

Client

Keyboard and mouse inputs

Capture GPU frame buffer

X / Gnome session

startx

TGX Sender

H.264 video w/ AAC audio

VM w/ vGPU

NVENC

CUDA
System management

- Systems are provisioned with a minimal kickstart file
- System configuration is managed with Ansible
  - CentOS and RHEL VM images
  - Hypervisor hosts
  - Virtualization manager (provisioning VMs)
  - (To-do) health checks and/or cleanup tasks on VMs
Testimonials

- Previously, I've been very limited in bandwidth here in the UK (often < 200 Kb), as I connect over SSH the StarCCM+ graphics front end to the CFD simulations running on the HPC. X Windows never worked and the connection has been painfully slow, with several minutes lag for several types of operation. TGX has overcome all these problems and is making the process of interacting with HPC more easy and efficient. – London, UK based CFD engineer

- Thanks for installing TGX on my laptop. I am glad that I did so as I am presently on vacation in India. And TGX has outperformed any other remote work option for me by a margin (VNC and EOD). The screen update is almost real time. – Houston, TX based geophysicist

- I love TGX, it has been a life saver for me and it is very stable. I have been on the same session and I just logout and back in and my work area is still there. – Houston, TX based geophysicist
NVIDIA vGPU Resources

NVIDIA Booth #1039 @ Red Hat Summit 2019

GTC 2019 Session:
S9299 – Deploying NVIDIA vGPU with Red Hat Virtualization (RHV)

Virtual GPU Test Drive
https://www.nvidia.com/tryvgpu

NVIDIA Virtual GPU Website
www.nvidia.com/virtualgpu

Questions? Ask on our Forums
https://gridforums.nvidia.com

NVIDIA Virtual GPU on LinkedIn
http://linkd.in/QG4A6u

Follow us on Twitter
@NVIDIAVirt

NVIDIA Virtual GPU YouTube Channel
http://tinyurl.com/gridvideos

#redhat #rhsummit
THANK YOU

plus.google.com/+RedHat
linkedin.com/company/red-hat
youtube.com/user/RedHatVideos
facebook.com/redhatinc
twitter.com/redhat