OPTIMIZING NVIDIA VIRTUAL GPU FOR THE BEST VDI USER EXPERIENCE

Erik Bohnhorst, NVIDIA
Agenda

- vGPU Introduction
- Virtual GPU October 2018 (vGPU 7.0)
- Architecting for Best User Experience
  - NVIDIA Recommended CPUs
  - What is the right GPU for your use case
  - NVIDIA vGPU Benchmarking
THE EVOLUTION OF MODERN WORKFLOWS

VISUAL WORKSPACE  MOBILITY  COLLABORATION  LARGE DATA INTERACTIVE HPC  AI  PHOTOREALISM  VR

VISUAL COMPUTING SPECTRUM
HOW IT WORKS

NVIDIA virtual GPU technology delivers a GPU experience to every desktop

With NVIDIA Virtual GPU
Driving the Best User Experience across simple to the most powerful Apps

- Apps and VMs
- NVIDIA Graphics Drivers
- NVIDIA Virtual GPU
- NVIDIA Virtualization Software
- Hypervisor
- NVIDIA Tesla GPU
- Server

CPU Only VDI
Limiting User Experience

- Apps and VMs
- Hypervisor
- Server

Server

Hypervisor

Apps and VMs
VIRTUAL GPU OCTOBER 2018 (vGPU 7.0)

Unprecedented Performance & Manageability

Multi-vGPU Support
World's Most Powerful Quadro vDWS

vMotion Support for vGPU
Live Migration of vGPU enabled VMs Quadro vDWS & GRID

NGC with vGPU
Available with vGPU Quadro vDWS

Tesla T4 GPU Support*
Latest Generation Turing Quadro vDWS

* Tesla T4 support coming with vGPU software 7.1 release
Greatest Leap Since 2006 CUDA GPU

**PASCAL**
- 11.8 Billion xtors
- 471 mm²
- 24 GB 10GHz

**TURING**
- 18.6 Billion xtors
- 754 mm²
- GDDR6 14GHz
Greatest Leap Since 2006 CUDA GPU

PASCAL

TURING

SHADER | COMPUTE
FP
or
INT

TENSOR CORE
FP16
INT8
INT4

RT CORE
Giga Rays/Sec

SHADER | COMPUTE
FP + INT
VIRTUAL GPU
TESLA T4 IS EXTREMELY VERSATILE

Enablement in Virtual GPU 7.1

- Great solution for
  - Quadro vDWS
  - GRID vPC
  - Deep Learning Inference
- 6 boards in high volume 2U rack servers

<table>
<thead>
<tr>
<th>TESLA T4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>1x TU104</td>
</tr>
<tr>
<td>Cores</td>
<td>2,560 CUDA Cores</td>
</tr>
<tr>
<td></td>
<td>320 Turing Tensor Cores</td>
</tr>
<tr>
<td></td>
<td>RT Cores</td>
</tr>
<tr>
<td>Memory</td>
<td>16 GB GDDR6</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Single Slot</td>
</tr>
<tr>
<td></td>
<td>(half height &amp; length)</td>
</tr>
<tr>
<td>Thermal</td>
<td>Passive</td>
</tr>
<tr>
<td>Power</td>
<td>70W - no external power</td>
</tr>
<tr>
<td>Max Users</td>
<td>16 (1GB FB)</td>
</tr>
<tr>
<td>Compute</td>
<td>65 FP16 TFLOPS</td>
</tr>
<tr>
<td></td>
<td>130 INT8 TOPS</td>
</tr>
<tr>
<td></td>
<td>240 INT4 TOPS</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>320 GB/s</td>
</tr>
</tbody>
</table>
94% FASTER RENDERING USING MULTI-GPU
SOLIDWORKS Visualize (Iray) Render Time

“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

Tests were run on a server with 2x Intel Xeon Gold (6154 3.0 GHz) CPUs, 512GB RAM, RHEL 7.5, NVIDIA Quadro vDWS software, Tesla V100-32Q, Driver - 410.39, 256 GB RAM, Windows 10 x64 RS3
Tests run on a sever with 2x Intel Xeon Skylake CPUs (Xeon 6148 2.4 GHz 32-core), NVIDIA Quadro vDWS software, Tesla V100 GPUs with 32Q profile, Driver - 410.53, 256 GB vRAM, Cent OS 7.4 64-bit. Benchmark Model: ~450-550 TFLOPs, 5.9M DOF, Highly Nonlinear Static, Axisymmetric model with non-axisymmetric loading and twist, Direct Sparse Solver (Model courtesy: SIMULIA)
NVIDIA VIRTUAL GPU SOFTWARE LINEUP

Quadro Virtual Data Center Workstation (Quadro vDWS)

For professional graphics applications; includes an NVIDIA Quadro driver.

Recommended GPU: Tesla P4*

GRID Virtual PC (GRID vPC)

For virtual desktops delivering standard PC applications, browser, and multimedia.

Recommended GPU: Tesla M10

GRID Virtual Applications (GRID vApps)

Use with VMware Horizon Apps.

Recommended GPU: Tesla M10

* P40 & V100 for High End & Ultra High-End Use Cases
* P6 for blade form factor deployments
NVIDIA RECOMMENDED CPU OPTIONS

Different workflows require different CPUs

GRID vPC
AMD EPYC CPU’s higher number of physical cores with lower frequency provide similar user experience to Intel Xeon Gold 6148 at higher scale.

Quadro vDWS
~3.0 GHz is required for many professional applications for optimal performance. Lower frequency can result in degraded performance.

GRID vPC

Intel Xeon Gold 6148
- 24 Cores @ 2.4 GHz

AMD EPYC 7501
- 32 Cores @ 2.0 GHz

Both CPUs provide similar user experience* while the AMD CPU can host ~25-33% more users

Quadro vDWS

Intel Xeon Gold 6154
- 18 Cores @ 3.0 GHz

Provides the required frequency per physical core and allows good scale (18 cores/CPU)

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames
**RECOMMENDED NVIDIA TESLA GPU OPTIONS**

Different workflows require different GPUs

### Quadro vDWS

| Intel Xeon Gold 6154 + | NVIDIA TESLA P4* + | Quadro vDWS |

### Grid vPC

| Intel Xeon Gold 6148 or AMD EPYC 7501 + | NVIDIA TESLA M10** + | Grid vPC |

### Grid vApps

| Intel Xeon Gold 6148 + | NVIDIA TESLA M10** + | Grid vApps |

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using SPECviewperf 12.1 on a Dell PowerEdge R740 with 2x Intel Xeon Gold 6154 CPUs

** Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames
QUADRO vDWS GUIDANCE

Deep learning, rendering, immersive visualization, and GPGPU compute applications

Largest CAD models, CAE, Photorealistic rendering, Seismic exploration, GPGPU compute

Large/complex CAD models, Seismic exploration, complex DCC effects, 3D Medical Imaging Recon

Large/complex CAD models, Advanced DCC, Medical Imaging

Medium size/complexity CAD models, Basic DCC, Medical Imaging, PLM

Small/simple CAD models, video, Entry PLM

Office, SketchUp | PACS/Diagnostics | Schlumberger, Halliburton, DeltaGen, Catia Live Rendering
AutoCAD, Revit, Inventor | | Ansys, Abaqus, Simulia
Solidworks, Siemens NX, Creo, Catia
Adobe CC Photoshop, Illustrator | Adobe CC Premiere Pro, After Effects, Autodesk Maya, 3ds Max, Mari, Nuke

Entry - Mid Range Quadro vDWS

High-End Quadro vDWS

Tesla P4 / T4

Tesla P40

Tesla V100
UP TO 6X TESLA P4

Best Density with 6x Tesla P4 on a Dual-Socket Server

NVIDIA recommends Intel Xeon Gold 6154 18-core 3.0 GHz which provides enough CPU resources to host 6x Tesla P4 GPUs with Quadro vDWS.

Tesla P4 benefits over Tesla M60:
- Performance*
- Price/Performance
- Smaller Form Factor
- Lower Power Consumption
- NVIDIA Pascal GPU Architecture Benefits

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using SPECviewperf 12.1 on a Dell PowerEdge R740 with 2x Intel Xeon Gold 6154 CPUs
TESLA P40/V100 FOR ULTRA HIGH END USERS

Tesla P40 and Tesla V100 power the most demanding workflows

Tesla P40 with Quadro vDWS for few high to ultra high end users.

Tesla V100 with Quadro vDWS for few high to ultra high end users and/or Deep Learning workflows.

When to choose Tesla P40 over P4:
- Maximum Performance*
- High Framebuffer profiles (12GB/24GB)

Multiple Tesla P4 GPUs are the most cost effective and flexible solution for many entry to mid range end users

REMEmBER:
Large framebuffer GPUs don’t guarantee high number of Quadro vDWS users

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using SPECviewperf 12.1 on a Dell PowerEdge R740 with 2x Intel Xeon Gold 6154 CPUs
SIMPLE vGPU LICENSE DECISION

Do you use CUDA or professional workstation apps?

- Yes! You need Quadro vDWS
  Includes vPC and vApps entitlement
- No

Do you use VDI? (Single user per OS)

- Yes! You need GRID vPC
  Includes vApps entitlement
- No

Do you have multiple users sharing a single OS through sessions? (RDSH, Horizon Apps, XenApp, etc.)

- Yes! You need vApps
**HOW NVIDIA MEASURES USER EXPERIENCE**

Applying Methodology of Physical PCs to Virtual PCs

**UX**

- **End User Latency**
- **Framerate**
- **Image Quality**
- **Functionality**

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>End User Latency</td>
<td>Measures Interactivity, how remote your session feels</td>
</tr>
<tr>
<td>Framerate</td>
<td>Measures the fluidity of your session</td>
</tr>
<tr>
<td>Image Quality</td>
<td>Measures the impact of the remote protocol</td>
</tr>
<tr>
<td>Functionality</td>
<td>Application and API compatibility</td>
</tr>
<tr>
<td>Consistency</td>
<td>Measures how consistent the UX is over time</td>
</tr>
</tbody>
</table>

**UNIQUELY** Quantifies Remote User Experience

**Metrics**

- **Host Resources**: CPU, GPU, Memory, etc.
- **Virtual Machine Resources**: vCPUs, vMemory, vGPU, IOPS, etc.
- **Network Consumption**: Bandwidth, etc.

+ **Monitors** Resource Utilization

= **Realistic** Sizing Recommendations
NVIDIA IMAGE QUALITY RECOMMENDATION

GRID vPC
YUV 4:4:4 for PC Users

Quadro vDWS
YUV 4:2:0 for Workstation Users

Reference Image

Reference Image

Reference Image

Reference Image
**YUV 4:4:4 IMPLICATIONS**

- **Improved Image Quality**
  - SSIM increase to 0.989**

- **Similar Bandwidth Utilization***
  - YUV 4:4:4 - 2% less bandwidth

- **Lower Remoted Frames***
  - YUV 4:4:4 - 9% fewer Remoted Frames

---

*Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload running 64 VMs with Tesla M10-1B

**Tested with NVIDIA Cirrus VDI Benchmarking tool and predefined reference images to represent multiple workflows
GRID vPC for Multiple Screens

1x 1080p Screen (CPU-Only)

End User Latency (ms)

Server CPU Utilization (%)

Remoted Frames / User

GRID vPC for High Screen Resolutions

1x 1080p Screen (CPU-Only)

End User Latency (ms)

Server CPU Utilization (%)

Remoted Frames / User

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames
GRID vPC for Multiple Screens

End User Latency (ms)

1x 1080p Screen (CPU-Only)
- Lower is better: 350
- Higher is better: 800

2x 1080p Screens (CPU-Only)
- Lower is better: 350
- Higher is better: 714

Server CPU Utilization (%)

Remoted Frames / User

GRID vPC for High Screen Resolutions

End User Latency (ms)

1x 1080p Screen (CPU-Only)
- Lower is better: 350
- Higher is better: 714

1x 4K Screens (CPU-Only)
- Lower is better: 350
- Higher is better: 714

Server CPU Utilization (%)

Remoted Frames / User

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames
GRID vPC for Multiple Screens

End User Latency (ms)

Server CPU Utilization (%)

Remoted Frames / User

GRID vPC for High Screen Resolutions

End User Latency (ms)

Server CPU Utilization (%)

Remoted Frames / User

* Tested with NVIDIA’s Cirrus VDI Benchmarking tool using the Knowledge Worker workload and comparing End-User Latency and Remoted Frames
THANK YOU