WHAT’S NEW IN GRID 7.0

Mason Wu, GRID & ProViz Solutions Architect
Nov. 2018
VIRTUAL GPU OCTOBER 2018 (GRID 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
DRIVING NEW WORKFLOWS
Empowering the Modern Digital Workplace

Windows 10
20% Increase in GPU

Photorealistic Rendering
Increasingly Complex Designs

Data Science
Increase in AI/DL & Inference
WHAT IS MULTI-vGPU?
Delivering a More Powerful Virtual Workstation

Multiple VM’s Sharing a GPU with NVIDIA Quadro vDWS

A Single VM Accessing Multiple Tesla GPUs with NVIDIA Quadro vDWS (7.0)
QUADRO vDWS MULTI-vGPU
Enabling New Workflows in Strategic Markets

Oil & Gas
Seismic interpretation, simulation

Manufacturing
Simulation, modeling & design

Federal Government
Simulation & training

Media & Entertainment
Rendering
SPEED RENDER TIME BY UP TO 4.3X

Blender Cycles (CUDA) Render Time

Render scenes up to 4.3X faster than CPU only with three Tesla V100-32Q GPUs assigned to a VM with NVIDIA Quadro vDWS.

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.
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
UP TO 6.8X FASTER SIMULATIONS
Ansys Fluent Computational Fluid Dynamics (CFD)

Internal Flow of a Cooling Water Jacket Model - 5.4 million cells
Ansys Fluent 19.1 with NVIDIA Quadro vDWS & Tesla V100-32Q

- 32 vCPU: Base + 2 HPC Pack Licenses
- 16 vCPU + 2x V100-32Q: Base + 14 HPC Licenses
- 16 vCPU: Base + 12 HPC Licenses
- 8 vCPU + 1x V100-32Q: Base + 5 HPC Licenses
- 8 vCPU: Base + 4 HPC Licenses
- 3 vCPU + 1x V100-32Q: Base License
- 4 vCPU: Base License

Tests were run on a server with 2x Intel Xeon Skylake (6148 2.4 GHz Turbo - 3.6 GHz), NVIDIA Quadro vDWS software, Tesla V100 GPUs with 32Q profile, Driver - 410.53, 256 GB RAM, Cent OS 7.4 64-bit. Benchmark Model: Water Jacket Model, Unsteady RANS, Internal Flow, Fluid- Water, size 4, 20 time steps.
UP TO 4.95X FASTER THAN CPU-ONLY

Abaqus/Standard 2018 Elastomeric Bearing Model

Abaqus with NVIDIA Quadro vDWS & Tesla V100-32Q

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)
vMOTION SUPPORT FOR vGPU

Delivering the Modern, Agile Data Center

Keep Servers Healthy

Run patches and upgrades at your own time, and easily perform workload balancing, without end user disruption or data loss.

Maximize Infrastructure ROI

Run mixed workloads using the same infrastructure by repurposing hosts that run VDI during the day to run HPC and other compute workloads at night.

vMotion supported on vSphere 6.7.1 and NVIDIA virtual GPU release 7.0

XenMotion support on Citrix XenServer 7.4 and NVIDIA virtual GPU release 6.0
ENABLING THE AGILE DATA CENTER

Ensure High Availability
Live migrate GPUs with support for vMotion & XenMotion

Lower IT Tickets
End-to-end monitoring with vROps integration

Maximize Infrastructure Utilization
Virtualize any workload and run VDI by day, HPC by night
ENABLING THE AGILE DATA CENTER
MIXED WORKLOADS WITH NVIDIA vGPU

Increase productivity & utilization, decrease costs
NVIDIA TESLA T4 FOR VIRTUALIZATION
Powering 3D Professional Virtual Workstations

<table>
<thead>
<tr>
<th>GPU Architecture</th>
<th>Turing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA Cores</td>
<td>2,560</td>
</tr>
<tr>
<td>Tensor Cores</td>
<td>320</td>
</tr>
<tr>
<td>Memory Size</td>
<td>16 GB</td>
</tr>
<tr>
<td>vGPU Profiles</td>
<td>1 GB, 2 GB, 4 GB, 8 GB, 16 GB</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 single slot</td>
</tr>
<tr>
<td>Power</td>
<td>70 W</td>
</tr>
<tr>
<td>Thermal</td>
<td>Passive</td>
</tr>
</tbody>
</table>
# NVIDIA TESLA GPUs

**Recommended for Virtualization**

<table>
<thead>
<tr>
<th>GPUs / Board (Architecture)</th>
<th>V100</th>
<th>P100</th>
<th>P40</th>
<th>P4</th>
<th>T4</th>
<th>M60</th>
<th>M10</th>
<th>M6</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Volta)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1 (Pascal)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>5,120</td>
<td>3,584</td>
<td>3,840</td>
<td>2,560</td>
<td>2,560</td>
<td>4,096</td>
<td>2,560</td>
<td>1,536</td>
<td>2,048</td>
</tr>
<tr>
<td>Memory Size</td>
<td>32 GB/16 GB HBM2</td>
<td>16 GB HBM2</td>
<td>24 GB GDDR5</td>
<td>8 GB GDDR5</td>
<td>16 GB GDDR5</td>
<td>16 GB GDDR5</td>
<td>32 GB GDDR5</td>
<td>8 GB GDDR5</td>
<td>16 GB GDDR5</td>
</tr>
<tr>
<td>vGPU Profiles</td>
<td>1 GB, 2 GB, 4 GB, 8 GB, 16 GB, 32 GB</td>
<td>1 GB, 2 GB, 4 GB, 8 GB, 16 GB</td>
<td>1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB</td>
<td>1 GB, 2 GB, 4 GB, 8 GB</td>
<td>1 GB, 2 GB, 4 GB, 8 GB, 16 GB</td>
<td>0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB</td>
<td>0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB</td>
<td>1 GB, 2 GB, 4 GB, 8 GB, 16 GB</td>
<td></td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Dual Slot &amp; SXM2 (rack servers)</td>
<td>PCIe 3.0 Dual Slot (rack servers)</td>
<td>PCIe 3.0 Dual Slot (rack servers)</td>
<td>PCIe 3.0 Single Slot (rack servers)</td>
<td>PCIe 3.0 Dual Slot (rack servers)</td>
<td>PCIe 3.0 Dual Slot (rack servers)</td>
<td>MXM (blade servers)</td>
<td>MXM (blade servers)</td>
<td></td>
</tr>
<tr>
<td>Power</td>
<td>250W/300W</td>
<td>250W</td>
<td>250W</td>
<td>75W</td>
<td>70W</td>
<td>300W (225W opt)</td>
<td>225W</td>
<td>100W (75W opt)</td>
<td>90W</td>
</tr>
<tr>
<td>Thermal</td>
<td>passive</td>
<td>passive</td>
<td>passive</td>
<td>passive</td>
<td>passive</td>
<td>active/passive</td>
<td>passive</td>
<td>bare board</td>
<td>bare board</td>
</tr>
</tbody>
</table>

**PERFORMANCE Optimized**

**DENSITY Optimized**

**BLADE Optimized**
Learn more about NVIDIA virtual GPU technology
www.nvidia.com/vgpu

Software Evaluation
https://www.nvidia.com/object/vgpu-evaluation.html

Test Drive NVIDIA virtual GPU today!
http://www.nvidia.com/testdrive
POP QUIZ

在GRID 7.1即将支援的GPU是?
在GRID 7的環境中，Deep learning開發者可以透過什麼讓Deep learning開發更方便快速?
GRID 7可以支援Multi-GPU的License type是？