Citrix XenApp / Microsoft RDSH – How to get the Best User Experience and Performance with NVIDIA vGPU Technology
G’day and Welcome

Jan Hendrik Meier
Citrix Systems GmbH
Senior Systems Engineer

@jhmeier

Thomas Remmlinger
NVIDIA
Senior GRID Solution Architect

@tremmlinger
Citrix XenApp und XenDesktop 7.15 LTSR

- Planning, Installation, Konfiguration and Administration
- For all Editions of XenApp and XenDesktop 7.15 LTSR
- Incl. Best Practices Troubleshooting
- only in German

GPU powered VDI – Virtual Desktops with NVIDIA GRID

- Available on Amazon
- 170 Pages about NVIDIA GRID
- Plan
- Implement
- Check
- Troubleshooting
- Now also in English!
| Founded 1851 |
| Headquarter Damme, Lower Saxony |
| ~ 2200 employee worldwide - ~1350 in Damme |
| 124 apprentices in 14 job types |
| Manufacturer for Potato, Sugar Beets und Vegetable Technique |
| World market leader in the Potato area |
vGPU Environment @ GRIMME

CAD Workload (Productive)
• 8 Servers
• Each Server with the following configuration:
  • 2 Xeon CPUs v4 @ 3.2 GHZ 8C / 16C HT
  • 768GB RAM
  • 2x NVIDIA M60
• Currently ~120 concurrent Users

XenApp / XenDesktop Workload (Productive)
• 4 Servers
• Each Server with the following configuration:
  • 2 Xeon CPUs v4 @ 3.6 GHZ 14C / 28C HT
  • 512GB RAM
  • 1x NVIDIA M10
• XenDesktop: Logistics and HR
vGPU Environment II @ GRIMME

- CAD (Test)
  - 3 Servers
  - Each Server with the following configuration:
    - 2 Xeon CPUs @ 3,5 GHZ 8C / 16C HT
    - 768GB RAM
    - 6x NVIDIA P4

- CAD Viewers (Productive)
  - 3 Servers
  - Each Server with the following configuration:
    - 2 Xeon CPUs @ 3,5 GHZ 8C / 16C HT
    - 768GB RAM
    - 6x NVIDIA P4
  - Currently ~70 concurrent Users
XenApp (v)GPU Tests
• Custom?
• Rasmus Raun-Nielsen Workload? (GTC Europe 2017 – vGPU for Office)
• LoginVSI?

Test Workload

LoginVSI + GPU Profiler Preview
(Big thanks to Jeremy for adding RDS Session measurement)
LoginVSI Workloads

- Task Worker
- Office Worker
- Knowledge Worker
- Power Worker
- AddOn: Multimedia Worker
- AddOn: Storage Workload
<table>
<thead>
<tr>
<th>Office Worker:</th>
<th>Multimedia:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe Reader</td>
<td>Adobe Acrobat</td>
</tr>
<tr>
<td>Internet Explorer</td>
<td>Internet Explorer</td>
</tr>
<tr>
<td>MS Excel</td>
<td>MS Excel</td>
</tr>
<tr>
<td>MS Outlook</td>
<td>MS Outlook</td>
</tr>
<tr>
<td>MS PowerPoint</td>
<td>MS PowerPoint</td>
</tr>
<tr>
<td>MS Word</td>
<td>MS Word</td>
</tr>
<tr>
<td>Photo Viewer</td>
<td>Google Chrome</td>
</tr>
<tr>
<td></td>
<td>Google Earth</td>
</tr>
<tr>
<td></td>
<td>HTML 5 3D Spinning Balls</td>
</tr>
<tr>
<td></td>
<td>MP3</td>
</tr>
<tr>
<td></td>
<td>Streaming video</td>
</tr>
</tbody>
</table>
## Test Environment

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dell 7910</td>
<td>XenServer 7.4</td>
</tr>
<tr>
<td>Intel Xeon E5-2687W v4</td>
<td>Windows Server 2016</td>
</tr>
<tr>
<td>64GB Ram</td>
<td>XenDesktop 7.15 CU1</td>
</tr>
<tr>
<td>Local SSDs</td>
<td>NVIDIA GRID 6.0 (390.42)</td>
</tr>
<tr>
<td></td>
<td>LoginVSI 4132</td>
</tr>
<tr>
<td></td>
<td>GPU Profiler (1.07 Preview)</td>
</tr>
</tbody>
</table>
Running Workload

Might be removed if we don’t get RAM for a Dell R730 to create the videos
LoginVSI Problems

FEHLMELDUNG GOOGLE EARTH SECOND USER!
LoginVSI Problems

The proxy server isn’t responding

- Check your proxy settings 127.0.0.1:8080.
  Go to Tools > Internet Options > Connections. If you are in Standard or Basic mode, you need to be in Advanced mode to change the proxy settings. If you are in Advanced mode, open the Advanced tab and select the Proxies button. In the Connect to the Internet using a proxy server section, select Use a proxy server and enter the IP address 127.0.0.1 and port 8080.
- Make sure your firewall settings aren’t blocking your connection.
- Ask your system administrator for help.
Three runs for each option to make sure there is no one time exception

**IMPORTANT:** During the tests only one VM was running on the Host / GPUs. If multiple VMs use the same Host and GPU results will change!
Results
Results – no vGPU

- CPU
- RAM
- GPU
- Frame Buffer
- Encoder
- Decoder
- Sessions
Results – P4-1A

CPU
RAM
GPU
Frame Buffer
Encoder
Decoder
Sessions
Results – P4-8A

CPU
RAM
GPU
Frame Buffer
Encoder
Decoder
Sessions
Frame Buffer
Framebuffer is at the limit at ~ 4 users. The framebuffer will begin swap to system memory.

GPU
Swapping the Framebuffer content to system memory will increase the GPU load.

CPU
High CPU load when the content is switching between Framebuffer and system memory.

Frame Buffer
Framebuffer usage goes up nearly proportional to the amount of users.

GPU
GPU load is going up with a higher amount of users but in a balanced way.. Depends on the applications.

CPU
CPU is already spiky... But a normal behaviour until the system have much things to handle in the logon process.
Results from the Nvidia Lab

<table>
<thead>
<tr>
<th>Number of Connections</th>
<th>Framebuffer Usage in MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>2161</td>
</tr>
<tr>
<td>11</td>
<td>2924</td>
</tr>
<tr>
<td>16</td>
<td>4403</td>
</tr>
<tr>
<td>19</td>
<td>5305</td>
</tr>
<tr>
<td>20</td>
<td>5476</td>
</tr>
<tr>
<td>23</td>
<td>5878</td>
</tr>
<tr>
<td>26</td>
<td>6489</td>
</tr>
<tr>
<td>27</td>
<td>6996</td>
</tr>
<tr>
<td>28</td>
<td>6988</td>
</tr>
<tr>
<td>29</td>
<td>7142</td>
</tr>
<tr>
<td>30</td>
<td>7584</td>
</tr>
<tr>
<td>31</td>
<td>7681</td>
</tr>
<tr>
<td>32</td>
<td>7988</td>
</tr>
</tbody>
</table>

1 XenApp screen per session => 1920 x 1200 resolution => Color Depth 24 => Defined by Citrix Policy

The framebuffer usage depends on the resolution and on the applications.
Virtualized applications are rendered in an off-screen buffer.

The „Offscreen Buffer“ handles the resolution for sessions and is responsible for the communication between session and GPU driver.

Multi-monitor and resolution support is different to VDI and limited by the virtualisation vendor e.g. VMware/Citrix.
Sizing Consideration
Results – P4-8A / P40-8A comparison

- CPU
- RAM
- GPU
- Frame Buffer
- Encoder
- Decoder
- Sessions
User Experience With / Without Graphics Card
Video with / without P4 – Browser / Youtube

Might be removed if we don’t get RAM for a Dell R730 to create the videos
Good to know
When using Internet Explorer as Published App or in shared hosted Desktop no GPU acceleration is available when using WS2008R2 or WS2012/WS2012R2 ➞ intended by Microsoft

No access to driver level features (no access to Nvidia controlpanel for User)

Not all features enabled or experiemental ( no CUDA or OpenCL with GRID 5.x, was experimental with GRID 4.x)

Hardware acceleration for WPF based Apps ( AutoCAD etc) need to be enabled through regkeys

Apps like Office 2010/2013/2016 have settings which may be disabled by admins for non-GPU usecases but apply to all users/servers

Avoid physical installation with more than 1 GPU for a XenApp VM running WS12R2/2016, it will not getting utilized

Good to know

Text areas fuzzy => Change Video Codec
Never forget
## Codecs

<table>
<thead>
<tr>
<th>BOARD</th>
<th>FAMILY</th>
<th>CHIP</th>
<th># OF CHIPS</th>
<th># OF NVDEC/CHIP</th>
<th>Total # of NDEC</th>
<th>H.264</th>
<th>H.265 (HEVC)</th>
<th>VP8</th>
<th>VP9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8 bit</td>
<td>10 bit</td>
<td>12 bit</td>
</tr>
<tr>
<td>TESLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tesla M10</td>
<td>Maxwell (1st Gen)</td>
<td>GM107</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Tesla M6</td>
<td>Maxwell (2nd Gen)</td>
<td>GM204</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Tesla M60</td>
<td>Maxwell (2nd Gen)</td>
<td>GM204</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Tesla P4 / P6</td>
<td>Pascal</td>
<td>GP104</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Tesla P40</td>
<td>Pascal</td>
<td>GP102</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Tesla P100</td>
<td>Pascal</td>
<td>GP100</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Virtualized applications are rendered in an off-screen buffer.

The „Offscreen Buffer“ handles the resolution for sessions and is responsible for the communication between session and GPU driver.

Multi-monitor and resolution support is different to VDI and limited by the virtualisation vendor e.g. VMware/Citrix.
NVENC for XenApp => Tom

Might be removed if we don’t get RAM for a Dell R730 to create the videos
THANK YOU