Real World Implementation (and snares) of a High Performance NVIDIA GRID Infrastructure
G‘day and Welcome

Jan Hendrik Meier
Grimme Landmaschinenfabrik
Insert st**** description here

@jhmeier

Thomas Remmlinger
NVIDIA
Senior GRID Solution Architect

@tremmlinger
Company - Grimme Landmaschinenfabrik GmbH & Co. KG

- 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
GPU powered VDI – Virtuelle Desktops with NVIDIA GRID

- Now available on Amazon 😊
- Currently only in German 😞 Sorry
- 170 Pages about NVIDIA GRID
- Plan / Implement / Check / Troubleshoot
Reasons for the Project

- New Product Lifecycle Management (including new CAD Software)
- Start: Training environment for 45 concurrent users necessary
  - Alternative: Rent / Buy Fat-Clients
- 67 Workstations with equal/less 8GB RAM
- Client age often more than 5 Years
Current GRID Environment

- **CAD Workload (Productive)**
  - 8 Servers
  - Each Server with the following configuration:
    - 2 Xeon CPUs @ 3.2 GHZ 8C / 16C HT
    - 768GB RAM
    - 2x NVIDIA M60
  - Currently ~150 concurrent Users

- **XenApp Workload (Not Productive yet)**
  - 4 Servers
  - Each Server with the following configuration:
    - 2 Xeon CPUs @ 3.6 GHZ 14C / 28C HT
    - 512GB RAM
    - 1x NVIDIA M10
Snare(s) when used as training environment
### Snare(s) when used as training environment

<table>
<thead>
<tr>
<th>Date Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fri Dec 02 2016 10:43:16</td>
<td>CPU 2 machine check error detected.</td>
</tr>
<tr>
<td>Fri Dec 02 2016 10:43:10</td>
<td>An OEM diagnostic event occurred.</td>
</tr>
<tr>
<td>Fri Dec 02 2016 10:43:10</td>
<td>A fatal error was detected on a component at bus 0 device 2 function 0.</td>
</tr>
<tr>
<td>Fri Dec 02 2016 10:43:10</td>
<td>A bus fatal error was detected on a component at slot 6.</td>
</tr>
</tbody>
</table>
Why did we continue with NVIDIA GRID?

• Stability fine after bug was fixed (function was disabled)
• Costs lower than with Fat-Clients (calculated for 5 years)
• More flexibility
• No CAD Offline Support any longer
  – At least required for Users in remote branches
  – Home Office
  – External Engineers
• Standardization – all Users have exactly the same installation
Planning Phase
Planning Phase

- Amount of Clients which needs to be replaced
  - 32 (2016)
  - 16 (2017)
- Cost Calculation
- Required amount of servers with Data Center redundancy
- Choosing the graphics card
- vGPU Profile
Cost Calculation

- Many variables
  - RAM necessary for each User
  - vGPU Profile => vGPU License
    (First year: GRID Virtual PC ~100$ vs GRID Quadro vDWS ~550$)
  - Amount of Users per GRID Card
  - Amount of Users per Server
- Requires a lot of Assumptions
Required amount of servers with Data Center redundancy

- 16 Users per Server (planned)
- 32 Users total
- Two redundancy options
  - Two Data Centers
    - 2 Servers per Data Center
    - 4 Servers total
  - Using three Data Centers
    - 1 Server per Data Center
    - 3 Servers total
Choosing the graphics card

• Which type of graphic card should I choose?
  – Identify your applications
    • What applications do you want to use?
    • What are the software vendor recommendations?
    • How old is your application and do you plan an update in the near future?

• Do a PoC
  • Do the PoC with different kind of user types
  • Nothing is more disturbing than months of implementation and nothing but angry users…
# NVIDIA TESLA GPUs

<table>
<thead>
<tr>
<th></th>
<th>M10</th>
<th>M60</th>
<th>P40</th>
<th>M6</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>4 NVIDIA Maxwell GPUs</td>
<td>2 NVIDIA Maxwell GPUs</td>
<td>1 NVIDIA Pascal GPU</td>
<td>1 NVIDIA Maxwell GPU</td>
<td>1 NVIDIA Pascal GPU</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>2,560 (640 per GPU)</td>
<td>4,096 (2,048 per GPU)</td>
<td>3,840</td>
<td>1,536</td>
<td>2,048</td>
</tr>
<tr>
<td>Memory Size</td>
<td>32 GB GDDR5 (8 GB per GPU)</td>
<td>16 GB GDDR5 (8 GB per GPU)</td>
<td>24 GB GDDR5</td>
<td>8 GB GDDR5</td>
<td>16 GB GDDR5</td>
</tr>
<tr>
<td>H.264 1080p30 streams</td>
<td>28</td>
<td>36</td>
<td>24</td>
<td>16</td>
<td>24</td>
</tr>
<tr>
<td>Max vGPU instances</td>
<td>64 (512 MB Profile)</td>
<td>32 (512 MB Profile)</td>
<td>24 (1 GB Profile)</td>
<td>16 (512 MB Profile)</td>
<td>16 (1 GB Profile)</td>
</tr>
<tr>
<td>vGPU Profiles</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, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 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>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Dual 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>
</tr>
<tr>
<td>Power</td>
<td>225W</td>
<td>240W / 300W (225W opt)</td>
<td>250 W</td>
<td>100W (75W opt)</td>
<td>90 W (70W opt)</td>
</tr>
<tr>
<td>Thermal</td>
<td>passive</td>
<td>active / passive</td>
<td>passive</td>
<td>bare board</td>
<td>bare board</td>
</tr>
</tbody>
</table>

**USER DENSITY**
- Optimized

**PERFORMANCE**
- Optimized

**BLADE**
- Optimized

---

Jan Hendrik Meier, Thomas Remmlinger
NVIDIA GTC EU 2017
Choosing the graphics card

- Only two options available (at the time of the project)
  - M10 (less powerful – more users)
  - M60 (more powerful – less users)

=> M60

- Now: more options available
  - Compare the details – don’t trust marketing suggestions (e.g. P4 is not suggested anywhere)
Choosing the vGPU Profile

- Monitor real world scenarios
  - Have a look at your users daily work
  - Find out their needs and bottlenecks
  - Use this data for a PoC
- Don’t use benchmark tools for sizing
- Think about future updates of your applications
- Don’t forget about memory, CPU and disk performance
Choosing the vGPU Profile

• Training:
  – Tests with CAD-Administrators, CAD-Vendor and IT
  – Two delivery groups (Purple and Violet)
• Productive
  – Selection of 3 possible vGPU-Types
  – 3 Key user Groups
  – Random access of Users to vGPU (without knowing which one is used)
Challenges
Challenges

- User acceptance
- CPU = Bottleneck?
- 3Dconnexion Devices
Challenges – User acceptance

• Sometimes loading is fast / sometimes slow
• Slow Performance
• Screen tearing
Challenges – Loading Slow
Challenges – Loading Slow
Challenges – Slow Performance

- Monitoring looked fine
- Session was visibly slow
- Only occurred on one host
- Solution
  - Bios Performance Setting was reset after an update
Challenges – Screen Tearing
Challenges – Screen Tearing

- Citrix Ticket
- Solution:
  - Ubuntu Problem
  - Minimization:
    - Modify / add \texttt{xorg.conf.d} for Intel / NVIDIA Graphics Card
    - Disable OpenGL Settings (Lighting, Texture Compression, Framebuffer object)
    - Disable all Effects
CPU = Bottleneck?

- Widely told: The CPU will be your main Bottleneck
- Reason: CAD Software is mostly still single core based
  - More vGPU‘s in a VM don‘t help
  - High CPU clock rate helps
- Tests with many Users
  - CPU = Bottleneck
- Production (our environment!)
  - CPU =! Bottleneck
Challenges – 3Dconnexion Devices
Challenges – 3Dconnexion Devices

• Standard Keyboard Keys on SpaceMouse (ESC, Alt,..) not working
• Reboot Message shown after logon
Challenges – 3Dconnexion Devices

• Standard Keyboard Keys not working
  – Standard Keyboard Keys require special registry key
    • *Customized functions for a 3Dconnexion SpaceMouse might not work in a VDA session.*
      [#LC4797] – Citrix VDA 7.11
    • Path: `HKLM\SYSTEM\CurrentControlSet\services\picakbf`
      Name: `Enable3DConnexionMouse`
      Type: `DWORD`
      Value: 1
Challenges – 3Dconnexion Devices

• Reboot message
  – Connect every(!) type of 3Dconnexion Device you have to your Master-Image (and reboot)
  – Although devices might look the same – they might be different revisions
Takeaways
Key Takeaways

- Monitoring
  - Historical Data!
- Tests which vGPU Profile is necessary
  - It’s not always the biggest necessary
- Don’t trust suggestions – test the possible Graphics cards and compare the prices
- Expect fast growing of the solution
- Hope someone blogged about not well documented registry keys ;}
Thank You!