SIZE DOES MATTER!

The ultimate sizing and optimization guide for GRID environments

Roy Textor, CEO – textor IT
Who am i?

• Roy Textor
• CEO @ textor IT – Kassel/Germany
• NGCA – NVIDIA GRID Community Advisor
• Citrix Solution Advisor / Citrix since over 20 Years
• Founder & Host of DCUG – German Citrix User Group
Agenda

1. Analysis
2. Hardware
3. Backbone
4. Design
5. GPU
6. Endpoints

SIZE DOES MATTER!
Agenda

1. Analysis
   - Requirements
   - Post-Monitoring

2. Hardware
   - Certified
     - BIOS
     - CPU
     - RAM
     - Disk
     - Hypervisor

3. Backbone
   - Network
     - Storage
     - Power/UPS

4. Design
   - OS
     - Deployment
     - Configuration
     - Applications

5. GPU
   - Which Board?
   - Which Profile?
   - Which License?

6. Endpoints
   - ThinClients
   - FatClients
   - Mobile Devices
1. Analysis

Pre-Analysis

• Start with existing workloads

• Record current work behavior

• Identify bottlenecks
1. Analysis

Pre-Analysis

• Start with existing workloads

• Record current work behavior

• Identify bottlenecks
1. Analysis

GPU Frequency of use

<table>
<thead>
<tr>
<th>Frequency of use (%)</th>
<th>GPU Workload (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>42%</td>
<td>1-5</td>
</tr>
<tr>
<td>13%</td>
<td>5-10</td>
</tr>
<tr>
<td>7%</td>
<td>10-15</td>
</tr>
<tr>
<td>5%</td>
<td>15-20</td>
</tr>
<tr>
<td>3%</td>
<td>20-25</td>
</tr>
<tr>
<td>3%</td>
<td>25-30</td>
</tr>
<tr>
<td>3%</td>
<td>30-35</td>
</tr>
<tr>
<td>2%</td>
<td>35-40</td>
</tr>
<tr>
<td>2%</td>
<td>40-45</td>
</tr>
<tr>
<td>2%</td>
<td>45-50</td>
</tr>
<tr>
<td>18%</td>
<td>50+</td>
</tr>
</tbody>
</table>
1. Analysis

![GPU FB Frequency of use](chart.png)

- **Frequency of use (%)**
  - 0% to 1%: 10%
  - 1% to 5%: 14%
  - 5% to 10%: 60%
  - 10% to 15%: 16%
  - 15% to 20%: 16%
  - 20% to 25%: 14%
  - 25% to 30%: 10%
  - 30% to 35%: 0%
  - 35% to 40%: 0%
  - 40% to 45%: 0%
  - 45% to 50%: 0%
  - 50% and above: 0%

**GPU FB Workload (%)**

**www.textor-it.de**
## 1. Analysis

<table>
<thead>
<tr>
<th></th>
<th>M60</th>
<th>Win7 x64</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profile - 2Q</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Memory (GB)</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPU Workload (%)</td>
<td>0%</td>
<td>77%</td>
</tr>
<tr>
<td>GPU FB (GB)</td>
<td>0.19</td>
<td>0.49</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPU Workload (%)</td>
<td>53%</td>
<td>36%</td>
</tr>
<tr>
<td>Memory (GB)</td>
<td>4.16</td>
<td>8.64</td>
</tr>
<tr>
<td><strong>Recommendation</strong></td>
<td></td>
<td>1GB FB</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12GB RAM</td>
</tr>
</tbody>
</table>
1. Analysis

**M60 / P40 – Quick compare**

<table>
<thead>
<tr>
<th></th>
<th>M60</th>
<th>P40</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPU</strong></td>
<td>2x Maxwell</td>
<td>1x Pascal</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>4096 (2048 per GPU)</td>
<td>3840</td>
</tr>
<tr>
<td><strong>Framebuffer</strong></td>
<td>16 GB (8 GB per GPU)</td>
<td>24 GB</td>
</tr>
<tr>
<td><strong>vGPU-Profiles</strong></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>
</tr>
</tbody>
</table>
1. Analysis

Requirements

• More GPU Power

• Less Framebuffer

• Similar CPU Power

• Less Memory
1. Analysis

Post-Monitoring

• Use available Monitoring tools/Suites
  - GPU Profiler 1.05 by Jeremy Main
  - Citrix Director
  - Lakeside Systrack
  - Control Up
  - Liquidware

• Compare with Pre-Analysis
2. Hardware

Server

- Only use certified servers
## 2. Hardware

### Server

- Only use certified servers

### Qualified Platform List (X10) for NVIDIA Tesla / NVIDIA GRID

<table>
<thead>
<tr>
<th>Server Models</th>
<th>Description</th>
<th>K40M*</th>
<th>K80</th>
<th>M40</th>
<th>M60</th>
<th>M10</th>
<th>P100</th>
<th>P40</th>
<th>P4</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBI-7128RG-X</td>
<td>2 NVIDIA Tesla (207U)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SBI-7128RG-F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBI-7128RG-F2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-1028GQ-TR</td>
<td>X10 1U 4 NVIDIA Kepler / Xeon Phi</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SYS-1028GQ-TRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-4028GR-TR</td>
<td>X10 4U 8 NVIDIA Kepler / Xeon Phi</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SYS-4028GR-TRT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-4028GR-TR2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-4028GR-TRT2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-1018GR-T</td>
<td>X10 1U 2 NVIDIA Kepler / Xeon Phi</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
<td>3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SYS-5018GR-T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-1028GR-TRT</td>
<td>X10 1U 3 NVIDIA Kepler / Xeon Phi</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SYS-1028GR-TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-2028GR-TRT</td>
<td>X10 2U 4 NVIDIA Kepler / Xeon Phi</td>
<td>✓</td>
<td>✓</td>
<td>1</td>
<td>1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SYS-2028GR-TR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-2028GR-TRHT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYS-2028GR-TRH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Hardware

Server

• Only use certified servers
• Check BIOS settings
2. Hardware

Server

- Only use certified servers
- Check BIOS settings
  (Above 4G Decoding / 64bit-MMIO / ... )
2. Hardware

Server

• Only use certified servers
• Check BIOS settings
• Check for CPU requirements
2. Hardware

Server

• Only use certified servers
• Check BIOS settings
• Check for CPU requirements
2. Hardware

Server

- Only use certified servers
- Check BIOS settings
- Check for CPU requirements
- Calculate Memory requirements
2. Hardware

Server

• Only use certified servers
• Check BIOS settings
• Check for CPU requirements
• Calculate Memory requirements
• Calculate Disk requirements (local)
2. Hardware

Server

- Only use certified servers
- Check BIOS settings
- Check for CPU requirements
- Calculate Memory requirements
- Calculate Disk requirements (local)
- Choose Hypervisor (Business case!)
3. Backbone

Network

• Bandwidth
3. Backbone

Network

- Bandwidth
- Number of hops
3. Backbone

Network

• Bandwidth
• Number of hops
• Latency
3. Backbone

Network

- Bandwidth
- Number of hops
- Latency
- Real throughput
3. Backbone

Storage

• Bandwidth
3. Backbone

Storage

• Bandwidth
• Number of hops
3. Backbone

Storage

• Bandwidth
• Number of hops
• Latency
3. Backbone

Storage

- Bandwidth
- Number of hops
- Latency
- Real throughput
4. Design

Operating System

• Windows 7
• Windows 10
• Windows Server 2012 R2
• Windows Server 2016
• Linux
4. Design

Deployment

• Native
• SCCM
• Citrix MCS / PVS
• ...
4. Design

Configuration

• Hardware default graphics adapter (Server OS)
4. Design

Configuration

• Hardware default graphics adapter (Server OS)
• Driver related optimization
4. Design

Configuration

- Hardware default graphics adapter (Server OS)
- Driver related optimization
- CATIA V5: Vertex Buffer Object
4. Design

Applications - Examples

• Google Chrome
4. Design

Applications - Examples

- Google Chrome
4. Design

Applications - Examples

• Google Chrome
• VLC
4. Design

Applications - Examples

• Google Chrome
• VLC
• MS Office
4. Design

Applications - Examples

• Google Chrome
• VLC
• MS Office
• Internet Explorer
5. GPU

Which Board?

M60 / P40 – Quick compare

<table>
<thead>
<tr>
<th></th>
<th>M60</th>
<th>P40</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>2x Maxwell</td>
<td>1x Pascal</td>
</tr>
<tr>
<td>Cores</td>
<td>4096 (2048 per GPU)</td>
<td>3840</td>
</tr>
<tr>
<td>Memory</td>
<td>16 GB (8 GB per GPU)</td>
<td>24 GB</td>
</tr>
<tr>
<td>vGPU-Profiles</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>
</tr>
</tbody>
</table>
## 5. GPU

### Which Board?

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU Architecture</th>
<th>CUDA Cores</th>
<th>Memory Size</th>
<th>H.264 1080p30 streams</th>
<th>Max vGPU instances</th>
<th>vGPU Profiles</th>
<th>Form Factor</th>
<th>Power</th>
<th>Thermal</th>
</tr>
</thead>
<tbody>
<tr>
<td>M60</td>
<td>NVIDIA Maxwell™</td>
<td>3,840</td>
<td>24 GB GDDR5</td>
<td>24</td>
<td>24 (1 GB Profile)</td>
<td>1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB</td>
<td>PCIe 3.0 Dual Slot (rack servers)</td>
<td>250 W</td>
<td>Passive</td>
</tr>
<tr>
<td>P40</td>
<td>1 NVIDIA Pascal GPU</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. GPU

Which Board?

NVIDIA TESLA P40 WITH NVIDIA QUADRO vDWS SOFTWARE DELIVERS UP TO 2X PERFORMANCE

<table>
<thead>
<tr>
<th>Software</th>
<th>NVIDIA® Tesla® M60-8Q</th>
<th>NVIDIA® Tesla® P40-24Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>3ds Max</td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>CATIA</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Creo</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Energy</td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Maya</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Medical</td>
<td>1.5</td>
<td>2.5</td>
</tr>
<tr>
<td>Showcase</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>Siemens NX</td>
<td>1.0</td>
<td>2.0</td>
</tr>
<tr>
<td>SOLIDWORKS</td>
<td>1.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Note: Comparing a single VM on NVIDIA Tesla M60-8Q vs a single VM on NVIDIA Tesla P40-24Q and based on SPECviewperf 12.1 benchmark.
5. GPU

Which Board?

M10

**SPECIFICATIONS**

<table>
<thead>
<tr>
<th>Specification</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtualization Use Case</td>
<td>Density-Optimized Graphics Virtualization</td>
</tr>
<tr>
<td>GPU Architecture</td>
<td>NVIDIA Maxwell™</td>
</tr>
<tr>
<td>GPUs per Board</td>
<td>4</td>
</tr>
<tr>
<td>Max User per Board</td>
<td>64 (16 per GPU)</td>
</tr>
<tr>
<td>NVIDIA CUDA® Cores</td>
<td>2560 NVIDIA CUDA Cores (640 per GPU)</td>
</tr>
<tr>
<td>GPU Memory</td>
<td>32 GB of GDDR5 Memory (8 per GPU)</td>
</tr>
<tr>
<td>H.264 1080p30 Streams</td>
<td>28</td>
</tr>
<tr>
<td>Max Power Consumption</td>
<td>225 W</td>
</tr>
<tr>
<td>Thermal Solution</td>
<td>Passive</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Dual Slot</td>
</tr>
</tbody>
</table>
5. GPU

Which Board?

TESLA M10 WITH NVIDIA GRID DELIVERS THE HIGHEST USER DENSITY FOR VIRTUAL GRAPHICS

<table>
<thead>
<tr>
<th>NVIDIA® GRID® K1</th>
<th>NVIDIA® Tesla® M10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

M10

SPECIFICATIONS

- **Virtualization Use Case**: Density-Optimized Graphics Virtualization
- **GPU Architecture**: NVIDIA Maxwell™
- **GPUs per Board**: 4
- **Max User per Board**: 64 (16 per GPU)
- **NVIDIA CUDA® Cores**: 2560 NVIDIA CUDA Cores (640 per GPU)
- **GPU Memory**: 32 GB of GDDR5 Memory (8 per GPU)
- **H.264 1080p30 Streams**: 28
- **Max Power Consumption**: 225 W
- **Thermal Solution**: Passive
- **Form Factor**: PCIe 3.0 Dual Slot
5. GPU

Which Profile?

<table>
<thead>
<tr>
<th>vGPU Profiles Supported</th>
<th>GRID Virtual Apps</th>
<th>GRID Virtual PC</th>
<th>Quadro Virtual Datacenter Workstation</th>
</tr>
</thead>
<tbody>
<tr>
<td>512 MB³</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1 GB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2 GB</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>3 GB⁴</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4 GB</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>6 GB⁴</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8 GB</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>12 GB⁴</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>16 GB⁵</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>24 GB⁴</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
5. GPU

Which License?

• **NVIDIA Quadro Virtual Datacenter Workstation**
  ○ For users who want to be able to use remote professional graphics applications with full performance on any device, anywhere.

• **NVIDIA GRID Virtual PC**
  ○ For users who want a virtual desktop but need great user experience leveraging PC Windows® applications, browsers, and high definition video.

• **NVIDIA GRID Virtual Applications**
  ○ For organizations deploying Citrix XenApp, VMware Horizon RDSH or other RDSH solution. Designed for PC level applications and server based desktops.
### 5. GPU

#### Which License?

<table>
<thead>
<tr>
<th>Feature</th>
<th>GRID Virtual Apps</th>
<th>GRID Virtual PC</th>
<th>Quadro Virtual Datacenter Workstation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>License Entitlement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concurrent User (CCU)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Capability Entitlement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Desktop Virtualization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDSH App Hosting</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Windows Guest OS</td>
<td>N/A</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Linux Guest OS</td>
<td>N/A</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Maximum Displays</td>
<td>N/A</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Maximum Resolution</td>
<td>N/A</td>
<td>2560*1600</td>
<td>4096*2160 (4K)</td>
</tr>
<tr>
<td>NVIDIA Quadro Software Feature</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUDA &amp; OpenCL Supported¹</td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>GPU Pass-through Supported¹</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>BareMetal Supported²</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>
5. GPU

Which License?
6. Endpoints

ThinClients

• Deep Compression decoding
• Min. 1.6 GHz CPU without H.264 offloading (>2.4 Ghz recommended)
• QuadCore CPU
• Multi Monitor Support
• Certified (HDX 3D Pro / Premium)
6. Endpoints

FatClients

- DXVA (Windows)
- VDPAU / VA-API (Linux)
6. Endpoints

Mobile Devices
Speed up...
Questions / Discussion?
TecCon 2017

• 13./14. November 2017
• Kassel / Germany
• Sessions / Workshops / Labs
• Ask-the-Expert
• Exhibition
• Community-Talks
• Party incl. Live-Act
• Food-Court

Win a Free Ticket at Booth E.39!

www.teccon.info
Contact me!

Exhibition:
Booth E.39

rt@textor-it.de
Thank You!