NVIDIA VIRTUALIZED GPU POWERS ANY AI WORKLOAD

John Fanelli, Vice President, NVIDIA GPU Virtualization
Anne Hecht, Senior Director, NVIDIA GPU Virtualization

December 18, 2019
AGENDA

Introduction

Latest vGPU Release - December 2019

NVIDIA vCompute Server

GRID vPC for Knowledge Workers

Quadro Virtual Workstation
VGPU EVERYWHERE FOR EVERYONE, EVERY WORKLOAD

GPU ACCELERATED DATA CENTER & CLOUD
EVOLUTION OF NVIDIA VIRTUAL GPU
NVIDIA VIRTUAL GPU SOFTWARE

GPU Virtualization for Every Workload

NVIDIA GRID Virtual PC/Applications
Office Productivity, Knowledge Worker VDI

Quadro Virtual Data Center Workstation
Performance Graphics

NVIDIA Virtual Compute Server
AI, Deep Learning, Data Science, HPC
NVIDIA VIRTUAL GPU

NVIDIA Virtual GPU Delivers GPU Acceleration to Every Visual and Compute Workload

With NVIDIA Virtual GPU

- Apps and VMs
- NVIDIA Graphics Driver
- NVIDIA Quadro Driver, or
- NVIDIA Compute Driver
- NVIDIA Virtual GPU
- NVIDIA virtualization software
- Hypervisor
- NVIDIA GPU
- Server

CPU Only Virtualization

- Apps and VMs
- Hypervisor
- Server
GTC CHINA 2019

GPU Virtualization for Every Workload

- **NVIDIA vGPU December 2019**
  - New Features, New GPUs Supported

- **NVIDIA Virtual Compute Server**
  - AI, Deep Learning, Data Science, HPC

- **China Leads in Cloud Adoption**
  - GPU Virtualization for Compute and Graphics

- Alibaba Cloud
- Tencent Cloud
NVIDIA VGPU DECEMBER 2019 (VGPU 10.0)
VIRTUAL GPU DECEMBER 2019 (10.0)  
Raising the Bar on User Experience, Performance & Operational Efficiency

Display Enhancements  
Up to 8K Resolutions

NVIDIA RTX 6000/8000  
High-Performance Graphics

NVIDIA V100S  
Most Advanced GPU for AI

Cross Branch Compatibility  
vGPU 10.0 and later
FLEXIBLE DISPLAY OPTIONS
Up to 8K Resolutions with Quadro vDWS

Support up to 8K (7680x4320) displays with NVIDIA Quadro vDWS powered virtual workstations¹

Support up to 5K (5120x2880) displays with NVIDIA Quadro vDWS and NVIDIA GRID powered virtual desktops²

Flexibly connect to any display from your virtual desktop

Move from desk to conference room to home and light up any display – up to 5K/8K resolution

¹ Maxwell GPUs will support up to 5K displays maximum.
² Windows 7 virtual machines support up to 4K displays maximum.
NVIDIA VIRTUAL GPU FLEXIBLE DISPLAY CONFIGURATIONS

Move your virtual machine to different endpoints and light up any display up to the highest number of pixels configured.

<table>
<thead>
<tr>
<th>Profile</th>
<th>Max Pixel Config</th>
<th>Max Pixel</th>
<th>1920x1080</th>
<th>5120x1440 curved</th>
<th>4096x2160 4K</th>
<th>5120x2880 5K</th>
<th>7680x4320 8K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1B</td>
<td>4x 2560x1600</td>
<td>16,384,000</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2B</td>
<td>2x 4K</td>
<td>17,694,720</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1Q</td>
<td>2x 4K</td>
<td>17,694,720</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2Q or larger</td>
<td>4x 4K</td>
<td>35,389,440</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4Q or larger</td>
<td>4x 5K</td>
<td>58,982,400</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>8Q or larger</td>
<td>2x 8K</td>
<td>66,355,200</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Assuming homogeneous display config
# NVIDIA DATA CENTER GPUs

## Recommended for Virtualization

<table>
<thead>
<tr>
<th>GPUs / Board (Architecture)</th>
<th>V100S / V100 NVLINK</th>
<th>RTX 8000 / RTX 6000</th>
<th>T4</th>
<th>M10</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volta</td>
<td>1 (Volta)</td>
<td>1 (Turing)</td>
<td>1 (Turing)</td>
<td>4 (Maxwell)</td>
<td>1 (Pascal)</td>
</tr>
<tr>
<td>Turing</td>
<td>1 (Turing)</td>
<td>4,608</td>
<td>2,560</td>
<td>2,560 (640 per GPU)</td>
<td>2,048</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>5,120</td>
<td>4,608</td>
<td>2,560</td>
<td>2,560 (640 per GPU)</td>
<td>2,048</td>
</tr>
<tr>
<td>Tensor Cores</td>
<td>640</td>
<td>576</td>
<td>320</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>RT Cores</td>
<td>---</td>
<td>72</td>
<td>40</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>FP64 TFLOPS</td>
<td>8.2 / 7.8</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>FP32 TFLOPS</td>
<td>16.4 / 15.7</td>
<td>14.9</td>
<td>8.1</td>
<td>---</td>
<td>6.2</td>
</tr>
<tr>
<td>Tensor TFLOPS</td>
<td>130 / 125</td>
<td>119</td>
<td>65</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Encoders HD FPS</td>
<td>1,080</td>
<td>630</td>
<td>510</td>
<td>840</td>
<td>720</td>
</tr>
<tr>
<td>Memory</td>
<td>32 GB/16 GB HBM2 1,132 GB/s</td>
<td>48 GB/24 GB GDDR6 624 GB/s</td>
<td>16 GB GDDR6 320 GB/s</td>
<td>32 GB GDDR5 (8 GB per GPU) 16 GB GDDR5 192 GB/s</td>
<td>MXM (blade servers)</td>
</tr>
<tr>
<td>Form Factor</td>
<td>PCIe 3.0 Dual Slot &amp; SXM2</td>
<td>PCIe 3.0 Dual Slot</td>
<td>PCIe 3.0 Single Slot</td>
<td>PCIe 3.0 Dual Slot</td>
<td>MXM (blade servers)</td>
</tr>
<tr>
<td>Power</td>
<td>250W/300W</td>
<td>250W</td>
<td>70W</td>
<td>225W</td>
<td>90W</td>
</tr>
<tr>
<td>Thermal</td>
<td>passive</td>
<td>passive</td>
<td>passive</td>
<td>passive</td>
<td>bare board</td>
</tr>
<tr>
<td>Use Case</td>
<td>Compute Performance</td>
<td>Graphics Performance</td>
<td>Perf/$$$</td>
<td>Cost Effective</td>
<td>Blade Optimized</td>
</tr>
</tbody>
</table>

- **Performance Optimized**
- **Density Optimized**
- **Blade Optimized**
RECOMMENDED NVIDIA GPU CONFIGURATIONS

- **Virtual GPU Software Edition**
  - **Use Case**: Creative & Technical Professional
    - **Compute Type**: Client Computing
    - **Virtual GPU Software Edition**: Quadro Virtual Data Center Workstation
    - **Recommended GPU**: NVIDIA T4 or Quadro RTX 6000, RTX 8000
  - **Use Case**: AI, Deep Learning, Data Science, & HPC
    - **Compute Type**: Server Workloads
    - **Virtual GPU Software Edition**: Virtual Compute Server
    - **Recommended GPU**: NVIDIA V100S, RTX 8000, RTX 6000, or T4
  - **Use Case**: Knowledge Worker
    - **Compute Type**: Client Computing
    - **Virtual GPU Software Edition**: GRID Virtual PC, GRID Virtual Apps
    - **Recommended GPU**: NVIDIA T4 or M10
RTX-POWERED VIRTUAL WORKSTATIONS
Quadro Anywhere, on Any Device

Extends the power of RTX platform to designers on any device, anywhere with Quadro Virtual Data Center Workstation (Quadro vDWS)

Flexibly provision virtual workstations and high-performance compute workloads from a single RTX Server with Quadro vDWS & NVIDIA vComputeServer

Take advantage of underutilized GPU resources with GRID vPC/vApps support for RTX 6000/8000

Now available from more OEM partners
**25X ACCELERATED RENDERING FOR NETFLIX**

NETFLIX Lost In Space scene: renders in a fraction of the time using RTX Server
- 6x faster for a single frame
- 25x faster for the entire shot

<table>
<thead>
<tr>
<th></th>
<th>CPU Node (Dual Skylake)</th>
<th>RTX Server (4 x RTX 8000)</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Render time (1 frame)</td>
<td>38 min</td>
<td>6 min</td>
<td>6x</td>
</tr>
<tr>
<td>Total render time (120 frames)</td>
<td>76 hours</td>
<td>3 hours</td>
<td>25x</td>
</tr>
<tr>
<td># of nodes</td>
<td>25</td>
<td>1</td>
<td>25x</td>
</tr>
<tr>
<td>Power (kW)</td>
<td>13.2</td>
<td>1.9</td>
<td>7x</td>
</tr>
<tr>
<td>Acquisition cost</td>
<td>$188k</td>
<td>$28k</td>
<td>7x</td>
</tr>
<tr>
<td>Cost of power (5 yrs.)</td>
<td>$68k</td>
<td>$10k</td>
<td>7x</td>
</tr>
<tr>
<td>Total cost</td>
<td>$256k</td>
<td>$38k</td>
<td>7x</td>
</tr>
</tbody>
</table>

*note: RTX Server not used in actual Lost In Space production*
Now Shipping: Arnold 6 with RTX Ray Tracing and AI Denoising

Up to 4.4x Performance over CPU rendering

Image courtesy of Lee Briggs, rendered in Arnold.
NVIDIA V100S FOR VIRTUALIZATION

Most Powerful GPU for AI

Power the most demanding AI, deep learning, data science and HPC workloads with NVIDIA vComputeServer

Take advantage of underutilized GPU resources and run Mixed Workloads

GRID vPC/vApps for knowledge worker VDI and

Quadro vDWS for graphics accelerated virtual workstations and simulation
V100S ~1.1X HPC & DL SPEEDUP

**Scientific Applications**

- **4x V100S over 4x V100 PCIe**
  - SPECFEM3D: 1.18x
  - AMBER: 1.06x
  - GROMACS: 1.08x

**Deep Learning Training Speedup**

- **1x V100S over 1x V100 PCIe**
  - ResNet-50 v1.5: 1.13x
  - BERT Large Fine Tuning: 1.06x

---

**SCIENTIFIC APPS CONFIG:**

<table>
<thead>
<tr>
<th>Application</th>
<th>Test Modules</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER [PME-Cellulose_NVE_4fs]</td>
<td>DC-Cellulose_NVE</td>
<td>18.17-AT_19.9</td>
</tr>
<tr>
<td>GROMACS [STMV - Dev Sandbox]</td>
<td>STMV - Dev Sandbox</td>
<td>gerrit_sandbox-puregpu_2019-08-15</td>
</tr>
<tr>
<td>SPECFEM3D</td>
<td>Four_material_simple_model</td>
<td>dvel_b7ed7a33</td>
</tr>
</tbody>
</table>

**DEEP LEARNING CONFIG:**

19.10 container, Mixed precision  |  ResNet-50 v1.5 MXNet, BS=256, ImageNet2012  |  BERT Large Fine Tuning Training TensorFlow, BS=10, SQuADv1.1
NEW - NVIDIA VIRTUAL COMPUTE SERVER
“AI IS THE MOST POWERFUL TECHNOLOGY FORCE OF OUR TIME”

Growth has outpaced a fading Moore’s Law, so increased demand for compute cannot be satisfied by simply upgrading the latest generation of server processors.

- 40% of large enterprises will use machine learning by 2023**
- 3M Data scientist and the most popular computer science major
- 75% of enterprises will implement a multi cloud-capable hybrid integration platform by 2021**

*Gartner. 2019 Strategic Roadmap for Compute Infrastructure. April 12 2019. ID G00375789
NEW NVIDIA vCOMPUTE SERVER

GPU Virtualization for Modern Workloads

All AI Workloads - DL Training & Inference, Data Science, HPC

GPU sharing and aggregation for improved utilization and affordability

GPU performance with hypervisor management tools, flexibility, and security

Supported by major hypervisor platforms VMware, Red Hat & Citrix
NVIDIA VIRTUAL COMPUTE SERVER
GPU Acceleration Features for Server Virtualization

New Features for vComputeServer

- Multi-VMs per GPU (Sharing)
- Multi-vGPU per VM (Aggregate)
- Peer-to-Peer over NVLink
- ECC & Page Retirement
- NVIDIA NGC (Containers)
- Enhanced, Flexible Scheduling

MANAGEMENT, MONITORING & MIGRATION
EASE DATA CENTER GPU ADOPTION

vComputeServer for Virtualized

GPU Performance & Reliability
For Virtualization

Maximize Utilization
Ensure right-size allocation with GPU sharing & aggregation

Management & Monitoring
Use same hypervisor virtualization tools to manager GPU Servers

Live Migration
Only vGPU solution to support live migration

Security
Hypervisor-based security extends to GPU workloads

Multi-Tenant
Isolate workloads to securely support multiple users
DEEP LEARNING TRAINING PERFORMANCE

Up to 50X Faster with NVIDIA vComputeServer & V100

Server Config: 2x Intel Xeon Gold (6140 3.2GHz) | VMware ESXi 6.7 U3, NVIDIA vComputeServer 9.1 RC, NVIDIA V100 (32C profile), Driver 430.18] TensorFlow Resnet-50 V1, NGC 19.01, FP16, BS:256

![Bar Chart]

Avg Images per Second
DEEP LEARNING INFERENCING PERFORMANCE

Up to 24X Faster with NVIDIA vComputeServer & T4

Avg Images per Second

vGPU T4

CPU only

Server Config: 2x Intel Xeon Gold (6140 3.2GHz) [VMware ESXi 6.7 U3, NVIDIA vComputeServer 9.1 RC, NVIDIA T4 (16C profile), Driver 430.43] TensorFlow Resnet-50 V1, NGC 19.01
COMPUTE, ANY WAY YOU WANT
With vComputeServer & NGC Ready Containers for Virtualization

NGC Containers on Bare Metal

NEW: GPU virtualization for compute
80-90% of server workloads are deployed on VMs*

NEW: NGC Containers for vSphere
70% of container deployments are on VMs*

CONTAINERS: SIMPLIFYING WORKFLOWS
For Bare Metal and Virtual GPU

NGC ON BARE METAL

- Most cost effective
- Maximum performance
- Static workloads
- When minimal manageability needed
- Good orchestration tool in place

NGC WITH vGPU

- Secure workload and user isolation
- Management and monitoring
- Multi-tenancy
- Mixed workloads
- Maximum GPU utilization
  - Fractional GPU
  - Flexible vGPU profile config
GRID VPC AND VAPPS
**WINDOWS 10**

Highest Graphics Requirement of any Operating System to Date

Windows XP

Windows 7

Windows 10

Windows 10 Requires a 50% Increase in graphics requirements, compared to Windows 7

*Percent of time consuming GPU (DirectX or OpenGL), comparing Windows 7 to Windows 10 builds in 2018 (builds 1803 and 1709)*
MODERN APPS IN THE DIGITAL WORKPLACE
Demands More Graphics

Applications that require graphics has doubled since 2012
Over half of enterprise users access at least one graphics accelerated app

Windows 7 vs Windows 10 (2017)
- 36% (↓)
- 53% (↓)
- 59% (↓)
- 64% (↓)
- 85% (↓)
- 409% (↓)

Windows 7 vs Windows 10 (2018)
- 49%
- 75%
- 66%
- 91%
- 98%
- 482%

*Percent of time consuming GPU (DirectX or OpenGL). Windows 10 builds in 2017 is based on data from builds 1709, 1703 & 1607. Windows 10 builds in 2018 is based on data from builds 1803 and 1709.
THE NEW DIGITAL WORKER
The Way We Work is Changing

Windows 10
50% increase in graphics usage over Windows 7

Office 365/Office 2016
50-100% increase in graphics usage over Windows 7

Web Browsers
Modern browsers are hardware accelerated by default

Latest Web Standards
Flash, HTML5, and WebGL are all very taxing to the CPU

PDF Viewers
Adobe® Acrobat® and Microsoft Edge are hardware accelerated by default

Collaboration and Video
Skype and YouTube are now prevalent across the enterprise

Digital Imaging & Design
Some features in Adobe® Photoshop® won’t work without a GPU

Multi-, High Res Monitors
Multi-monitors is the new normal and 4K is becoming mainstream

1 Percent of time consuming GPU (DirectX or OpenGL), comparing Windows 7 to Windows 10 builds in 2018 (builds 1803 and 1709)
# NVIDIA DATA CENTER GPUS
Empowering the Modern Digital Workplace

<table>
<thead>
<tr>
<th></th>
<th>2 x NVIDIA T4</th>
<th>1 x NVIDIA M10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>32 users</td>
<td>32 users</td>
</tr>
<tr>
<td>Form Factor</td>
<td>2X PCIe 3.0 single slot</td>
<td>PCIe 3.0 dual slot</td>
</tr>
<tr>
<td>Power</td>
<td>140W</td>
<td>225W</td>
</tr>
<tr>
<td>CODECs</td>
<td>VP9, H.265</td>
<td>H.264</td>
</tr>
<tr>
<td>System Memory Support</td>
<td>&gt; 1TB</td>
<td>&lt; 1TB</td>
</tr>
<tr>
<td>Use Case</td>
<td>Universal GPU for virtual workstations, knowledge workers, rendering, inferencing, training</td>
<td>Lowest TCO for knowledge workers</td>
</tr>
</tbody>
</table>
NVIDIA T4 FOR VIRTUAL PCs
Optimize Data Center Utilization with Mixed Workloads

• Two NVIDIA T4 GPUs support the same user density as a single M10 and fit in the same 2 slot PCIe form factor.

** Tested by NVIDIA on knowledge worker workloads (Excel, Word, PowerPoint, Chrome, Media Player, PDF) running on a single HD-resolution display with NVIDIA GRID 7.1 and NVIDIA T4-1B

33% BETTER PERFORMANCE WITH NVIDIA GRID
UX BASED ON REMOTED FRAMES
NVIDIA GRID vPC FOR THE BEST WINDOWS 10 VDI USER EXPERIENCE

- Up to 15% Improved End User Latency*
- Uncompromised Image Quality
- Up to 25% More Remoted Frames*
- Up to 60% Lower CPU Utilization*

* Improvement due to NVIDIA GRID vPC will be dependent on configuration and workload. Tested by NVIDIA on knowledge worker workloads (Excel, Word, PowerPoint, Chrome, PDF) running single 4K resolution monitor on VMware Horizon 7.9, Windows 10 1803, Office 2016 with NVIDIA T4-2B and NVIDIA virtual GPU 9.0 release, with 32 users per GPU.
SAFEGUARD ADOPTION WITH NVIDIA GRID

With NVIDIA GRID, police officers can use Google Earth and write reports from a virtual session, from anywhere. Simplified management enabled IT to expand VDI to knowledge workers and city planning engineers with 75% leaner IT compared to a city with the same population.
Häagen-Dazs’ Modern WebGL enabled Website Boosts Its Brand Image

“When we compared before and after introducing NVIDIA virtual GPUs, performance improvement was significant. With NVIDIA, it was obvious that the display and rendering of content was a lot faster.... Today, our VDI system performs so well that we simply don’t hear from our users.”

Shinichi Takeshita
Manager of Information Systems, Häagen-Dazs Japan
QUADRO VIRTUAL WORKSTATION
THE MOST POWERFUL VIRTUAL WORKSTATION
Quadro vDWS accelerates & streamlines professional workflows anywhere

Manufacturing | Architecture | Medical Imaging | Energy

Media and Entertainment | Education | Federal | Transportation
NVIDIA T4 FOR VIRTUALIZATION
NVIDIA Turing Generation of Computer Graphics on a Quadro Virtual Workstation

Virtual Quadro Workstation for the Professional Designer & Data Scientist

• Up to 2X graphics performance versus M60
• 5 Giga Rays per second for real-time, interactive rendering
• NGC support; run deep learning inferencing workloads 25x faster than CPU on a virtual machine
LARGER PROFILES BOOST PERFORMANCE

NVIDIA Quadro

Measured using SPECviewperf 13 benchmark comparing Geomean running NVIDIA GRID vPC with T4-2B and NVIDIA Quadro vDWS with T4 2Q, 4Q, and 16Q profiles.
More than 9,000 VDI systems take advantage of improved performance and UX

Key Applications: Dassault CATIA
QUADRO RTX FOR VIRTUAL WORKSTATIONS

Key Features:

- Quadro RTX 8000 or RTX 6000 GPUs
- RTX Server with flexible deployment with NVIDIA Quadro vDWS
- Boost desktop application rendering
- Up to 96 GB of ultra-fast memory with NVLink
LATEST GENERATION QUADRO VIRTUAL WORKSTATION PERFORMANCE

Work Faster with Larger Models

1.4x improved performance with Quadro RTX 6000/8000 for virtual workstations

Greater performance/$ with Quadro RTX 6000/8000 for virtual workstations

Added AI support and ray tracing support with Tensor and RT cores

RTX 3D Graphics: 1.4x performance
SPECviewperf13 Geomean

Quadro Virtual Workstations
A GPU FOR EVERY VIRTUAL WORKLOAD

**NVIDIA M10**
Knowledge Worker VDI w/NVIDIA GRID

**NVIDIA T4**
Entry — Mid Range Quadro vDWS
Inference w/Virtual ComputeServer
Knowledge Worker VDI w/NVIDIA GRID

**NVIDIA RTX 6000 and 8000**
High-End Quadro vDWS

**NVIDIA V100s**
High-End Virtual Compute

**Office Productivity, streaming video**

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

**Enterprise Acceleration, Graphics, Analytics, Inference**

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

**Largest CAD models, CAE, Photorealistic rendering, Seismic exploration**

**Deep Learning Training, HPC, AI, Data Science**
GTC CHINA
VGPU DEMOS HOSTED BY PARTNERS
VGPU SESSIONS

Virtual GPU powers Ali cloud computing service [CN9241]

Lecturer: Gao Feng, Technical Expert, Alibaba Cloud Technology Co., Ltd.
2019/12/18 Wednesday 13:50-14:35 | Session 4 (104 105)

GPU virtualization brings value to IaaS [CN9832]

Lecturer: Gao Ping, Tencent Expert Product Manager
2019/12/19 Thursday 09:50-10:35 | Session 4 (104 105)

New optimization and improvement of NVIDIA vGPU in Linux KVM [CN9309]

Lecturer: Neo Jia, Chief Architect, vGPU, NVIDIA Corporation
2019/12/19 Thursday 13:15-14:00 | Session 4 (104 105)
NVIDIA VIRTUAL GPU RESOURCES

Virtual GPU Test Drive
https://www.nvidia.com/tryvgpu

NVIDIA Virtual GPU Website
www.nvidia.com/virtualgpu

NVIDIA Virtual GPU YouTube Channel
http://tinyurl.com/gridvideos

Questions? Ask on our Forums
https://gridforums.nvidia.com

NVIDIA Virtual GPU on LinkedIn
http://linkd.in/QG4A6u

Follow us on Twitter
@NVIDIAVirt
**SELECTING THE RIGHT PROFILE SIZES**

**Typical Customer Deployment Scenario**

<table>
<thead>
<tr>
<th>Knowledge Worker</th>
<th>Light User</th>
<th>Medium User</th>
<th>Heavy User</th>
<th>Compute Workloads</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>32 - 96 Users per Server</strong></td>
<td><strong>16 - 24 Users per Server</strong></td>
<td><strong>12-18 Users per Server</strong></td>
<td><strong>6-9 Users per Server</strong></td>
<td><strong>Virtual Workstation</strong></td>
</tr>
<tr>
<td>GPU: T4 or M10</td>
<td>GPU: T4</td>
<td>GPU: T4</td>
<td>GPU: RTX 6000 or 8000</td>
<td>GPU: V100S or RTX 8000, RTX 6000 or T4</td>
</tr>
<tr>
<td>GPUs/server: 4-6 T4 or 2-3 M10</td>
<td>GPUs/server: 4-6</td>
<td>GPUs/server: 4-6</td>
<td>GPUs/server: 2-3</td>
<td>GPUs/server: 4-6 T4 or 2-3 V100S or RTX</td>
</tr>
<tr>
<td>Profiles: T4-1B or 2B, M10 1B or 2B</td>
<td>Profiles: T4-2Q</td>
<td>Profiles: T4-2Q/T4-4Q</td>
<td>Profiles: RTX6000-8Q/RTX6000-12Q</td>
<td>CPU: Intel 6242</td>
</tr>
<tr>
<td>CPU: Intel 6254, 4vCPU</td>
<td>CPU: Intel 6254, 8vCPU</td>
<td>CPU: Intel 6254, 12vCPU</td>
<td>CPU: Intel 6254</td>
<td></td>
</tr>
<tr>
<td>Memory: 8-16GB RAM</td>
<td>Memory: 16-32GB RAM</td>
<td>Memory: &gt;96GB RAM</td>
<td>Memory: &gt;96GB RAM</td>
<td></td>
</tr>
<tr>
<td>CPU: Intel 6254</td>
<td>CPU: Intel 6254</td>
<td>CPU: Intel 6254</td>
<td>CPU: Intel 6254</td>
<td></td>
</tr>
</tbody>
</table>

Virtual Desktop: NVIDIA GRID vPC/vApps

Virtual Workstation: NVIDIA Quadro vDWS

Guidance based on actual customer data, actual performance may vary.
# NVIDIA VIRTUALIZED DATA CENTER

New vComputeServer Enables vGPU Across the NVIDIA Platform

## CUSTOMER USE CASES
- Speech
- Translate
- Recommender
- Healthcare
- Manufacturing
- Finance
- Molecular Simulations
- Weather Forecasting
- Seismic Mapping
- Creative & Technical
- Knowledge Workers

## CONSUMER INTERNET & INDUSTRY APPLICATIONS

## SCIENTIFIC APPLICATIONS
- Amber
- NAMD
- +600 Applications

## PRO VIZ & GRAPHICS

## APPS & FRAMEWORKS
- python™
- TensorFlow
- mxnet
- Chainer
- ONNX

## CUDA-X & NVIDIA SDKs
- MACHINE LEARNING
  - cuDF
  - cuML
  - cuGRAPH
- DEEP LEARNING
  - cuDNN
  - CUTLASS
  - TensorRT
- HPC
  - OpenACC
  - cuFFT

## VISUALIZATION
- OPTIX
- NVEncode
- NVDecode

## CUDA & CORE LIBRARIES - cuBLAS | NCCL

## VIRTUALIZATION
- Virtual Compute Server

## vGPU

## NVIDIA GPUs & SYSTEMS
- NVIDIA GPU
- RTX GPU
- SYSTEM OEM

## CLOUD
- NVIDIA GPU SYSTEM OEM
- vDWS
- GRID vPC
- GRID vApps

## NVIDIA VIRTUALIZED DATA CENTER

## CUDA & CORE LIBRARIES - cuBLAS | NCCL

## VIRTUALIZATION
- Virtual Compute Server

## vGPU

## NVIDIA GPUs & SYSTEMS
- NVIDIA GPU
- RTX GPU
- SYSTEM OEM

## CLOUD
- NVIDIA GPU SYSTEM OEM
- vDWS
- GRID vPC
- GRID vApps
NEW NVIDIA LICENSING PORTAL
Enhanced User Experience for Licensing

✓ Entitlement Dashboard
✓ Intuitive License Generation
✓ Streamlined Software Downloads
✓ Simplified License Server Management
Learn more about NVIDIA virtual GPU technology
www.nvidia.com/virtualgpu

Find an NVIDIA virtual GPU certified server
https://www.nvidia.com/object/vgpu-certified-servers.html

Read our Virtual GPU Sizing Guides

Find NVIDIA vGPU Documentation
https://docs.nvidia.com

Try NVIDIA virtual GPU today!

See us at GTC San Jose in March!
VCOMPUTESERVER BRINGS NVIDIA NGC TO VSPHERE

Extend your NGC-Ready Systems & NGC Support Services to vSphere

Registry of GPU Accelerated Software
Accelerate AI to production with 150+ NGC Containers, Pre-trained models, Training scripts and workflows in vSphere

vSphere to manage NGC containers in VMs
Use vSphere to streamline management of VMs running NGC AI containers on GPUs

Validated Servers & Enterprise-grade Support
Performance validated GPU servers for faster rollout & direct access to NVIDIA's experts for support with NGC software
NGC CONTAINERS: ACCELERATING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments
• Eliminates complex, time-consuming builds and installs

Get started in minutes
• Simply Pull & Run the app

Portable
• Deploy across various environments, from test to production with minimal changes

WHY NGC CONTAINERS

Optimized for Performance
• Monthly DL container releases offer latest features and superior performance on NVIDIA GPUs

Scalable Performance
• Supports multi-GPU & multi-node systems for scale-up & scale-out environments

Designed for Enterprise & HPC environments
• Supports Docker & Singularity runtimes
• Support for VMware vSphere

Run Anywhere
• NVIDIA Pascal/Volta/Turing-powered GPUs in PCs, workstations, and servers
• From Core to the Edge
• On-Prem to Hybrid to Cloud
• On bare metal or in a virtual machine
DEEP LEARNING INFERENCING PERFORMANCE

Up to 25X Faster with NVIDIA vComputeServer & T4

Server Config: 2x Intel Xeon Gold (6140 3.2GHz) [VMware ESXi 6.7 U3, NVIDIA vComputeServer 9.1 RC, NVIDIA T4 (16C profile), Driver 430.43] TensorFlow Resnet-50 V1, NGC 19.01
TREATING PHYSICAL AND VIRTUAL PCs THE SAME

Applying Known Metrics from Physical PCs to Virtual PCs

**END USER LATENCY**
Measures the end user responsiveness

**FRAMERATE**
Measures the fluidity of your session

**IMAGE QUALITY**
Measures the impact of the remote protocol

*User observes change*  
*User input*  

Patent Filing in Progress
NVIDIA GRID FOR LOWER CPU UTILIZATION

Compared to CPU only VDI environments, NVIDIA GRID provides:

• Up to 60% lower average CPU utilization
• Well below 80% CPU utilization with a lot of headroom
• Less spikes in CPU utilization

Tested by NVIDIA on knowledge worker workloads (Excel, Word, PowerPoint, Chrome browsing with WebGL, Media Player, PDF) running single 4K resolution monitor on VMware Horizon 7.9 on vSphere ESXi 6.7 u3, Windows 10 1803, Office 2016 with NVIDIA Tesla T4-2B and NVIDIA virtual GPU 9.0 release.