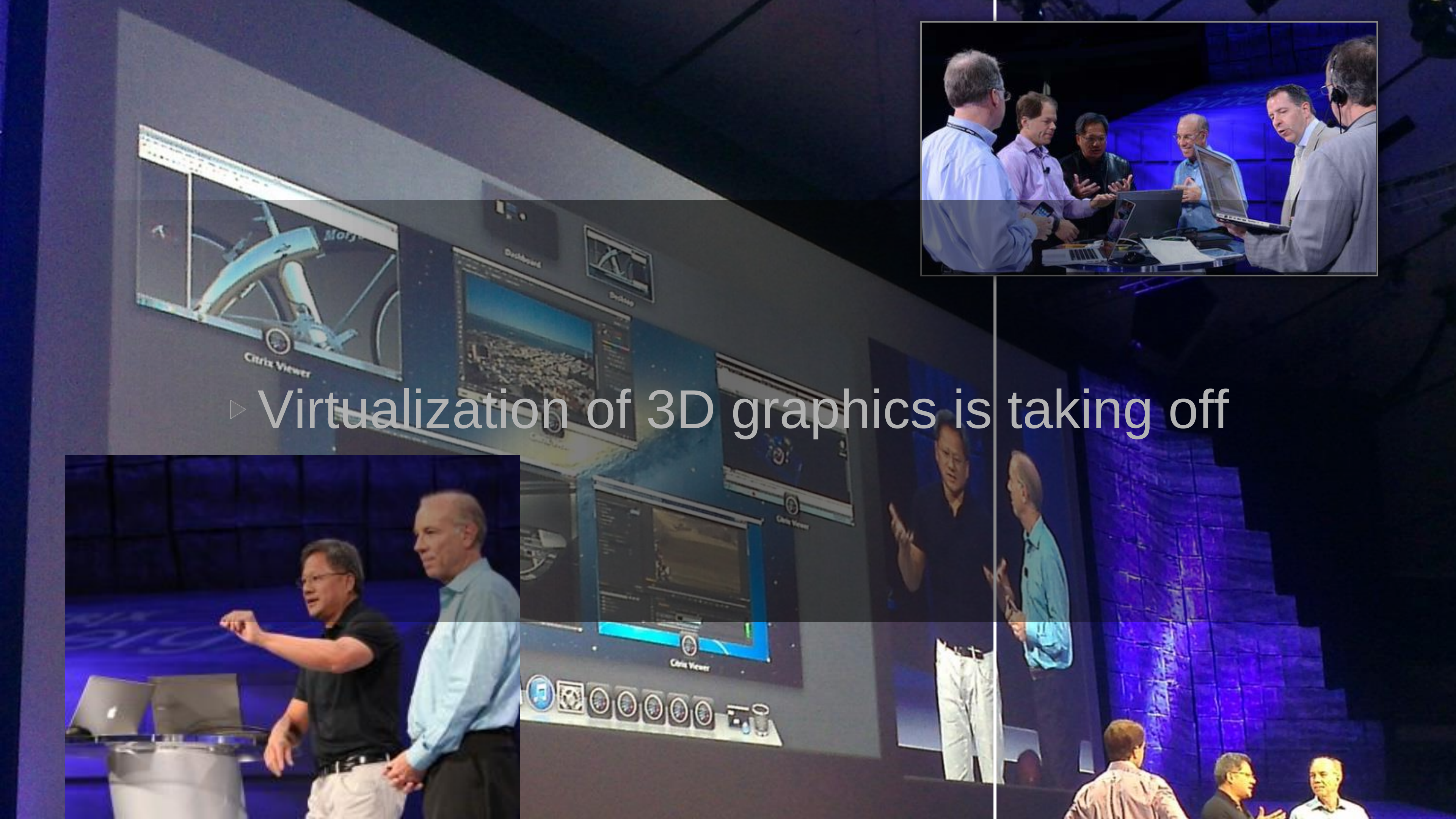
GPU TECHNOLOGY WORKSHOP SOUTH EAST ASIA 2014

Delivering virtualized 3D graphics apps with Citrix XenDesktop & NVIDIA Grid GPUs

Garry Soriano
Solution Engineer, ASEAN
Citrix Systems

garry.soriano@citrix.com

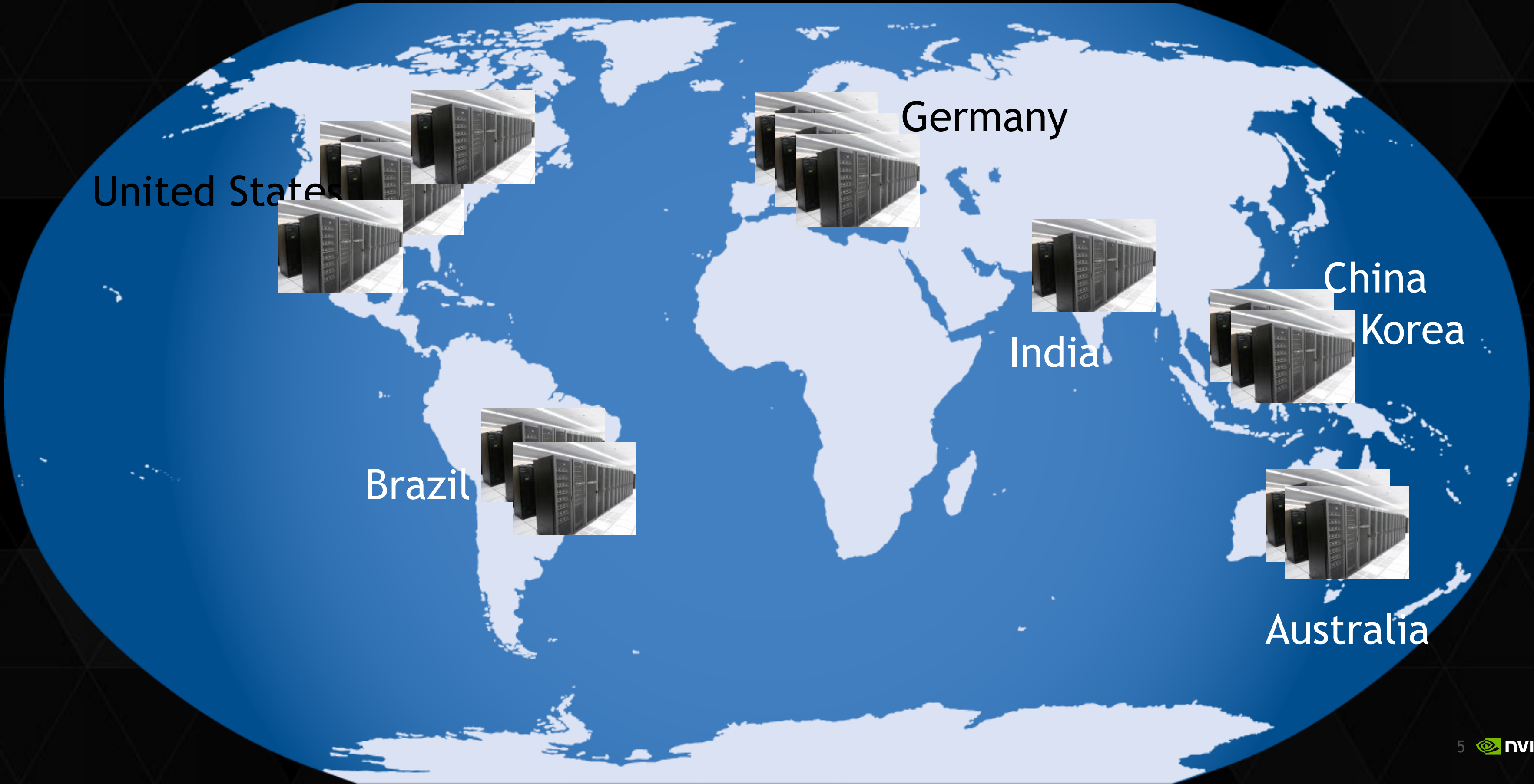


- Virtualization of 3D graphics is taking off

POWERFUL AND WIDE RANGE OF BUSINESS DRIVERS

- Global workforce
- Security of intellectual property
- Time-to-market
- Work from anywhere
- Disaster recovery
- Cost efficiencies

GLOBAL PRODUCT DEVELOPMENT TEAMS - REAL EXAMPLE



GLOBAL DEVELOPMENT EFFORT – REAL EXAMPLE

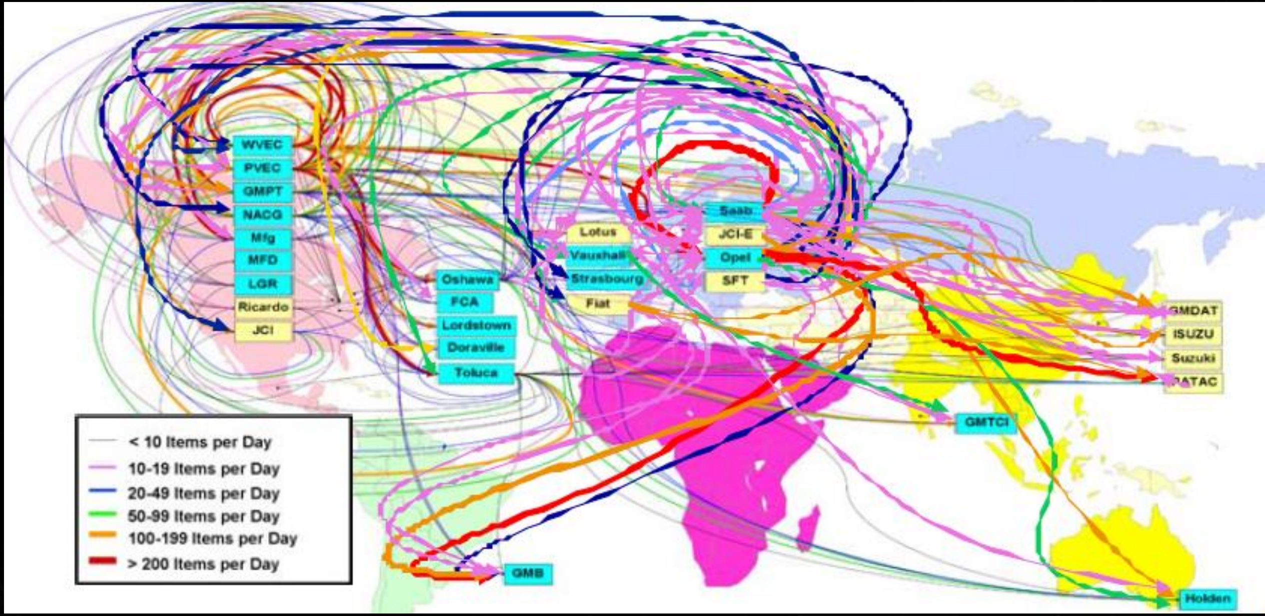
- 30,000 CAD files or 70 GB of data to be synchronized every day

Across 26 design centers (30,000+ users)

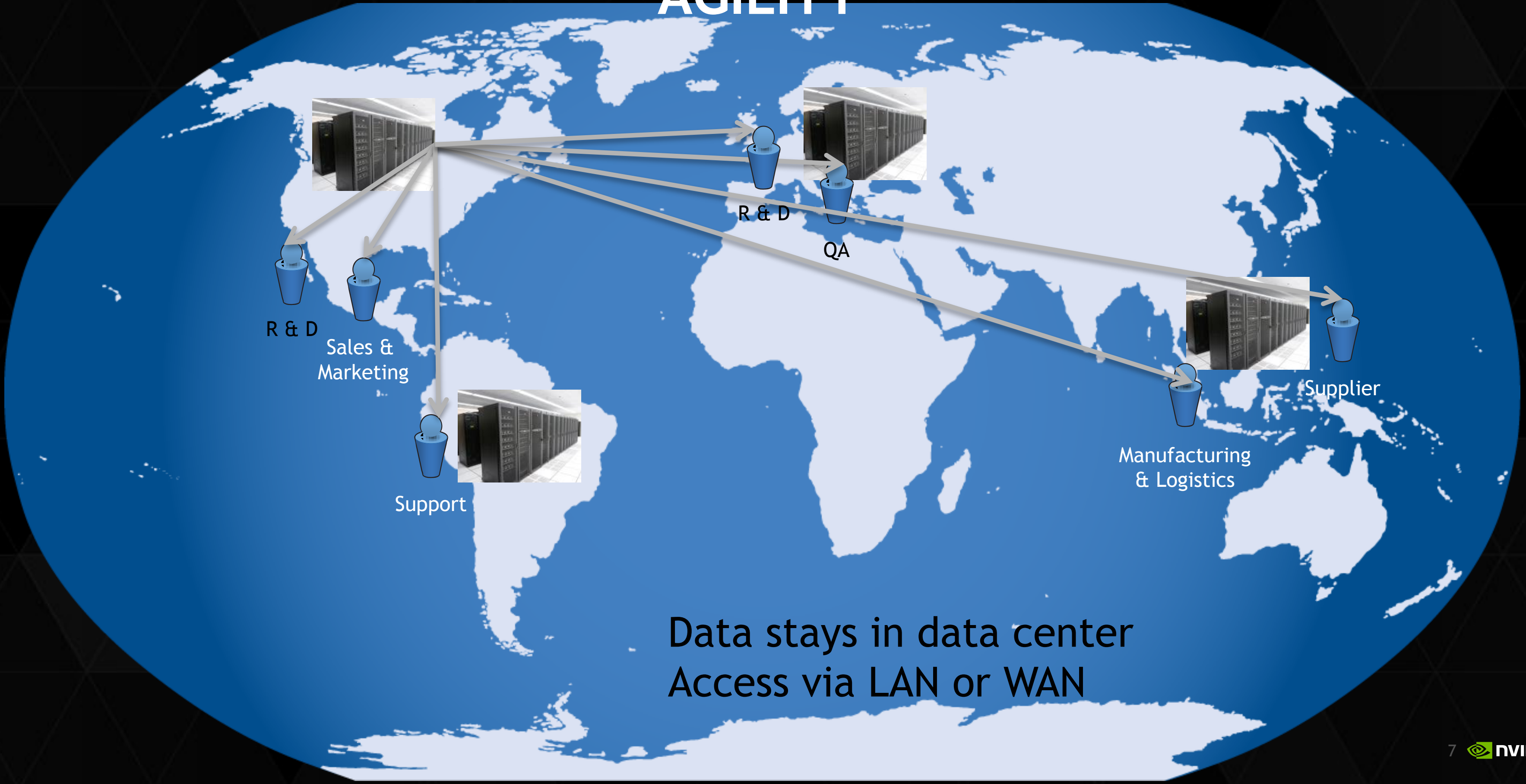
Across 16 countries

It took 2 weekends to sync all code updates!

More challenging for 4,000+ suppliers and partners



ENHANCED IP CONTROL, COLLABORATION AND GLOBAL AGILITY



Data stays in data center
Access via LAN or WAN

CITRIX CUSTOMERS USING GPU ACCELERATION

ABB

HANSA

Vestas®

BOMBARDIER

VOLVO

DAIMLER

Gulfstream™

JOHN DEERE

RAMBØLL

Wiha®

RaySearch Laboratories

BRÜCKNER GROUP

KNIGHTEC

MAERSK OIL

IMSCAD™

KPF

INTENSIV FILTER HIMENVIRO™
Environmental Technologies
WE CARE FOR AIR

VIRTUALIZE GRAPHICS WORKSTATIONS IN THE CLOUD

HDX 3D Pro Clients

XenApp

Windows app virtualization

Mobilize Windows apps for maximum security, control and performance

XenDesktop

Windows app and desktop virtualization

Deliver virtual Windows desktops with the best cost, performance and security for every business need

HDX 3D Pro

Deliver desktops and apps with best performance using GPU acceleration

XenServer

Open source platform for cost-effective cloud, server, and desktop virtualization infrastructures

COMPONENTS FOR HDX 3D PRO

www.citrix.com/go/vgpu

- Shared GPU for Desktops
  - XenDesktop 7.5
  - XenServer 6.2 Service Pack 1
  - NVIDIA GRID K1 and K2 boards
  - Latest NVIDIA GRID vGPU **Drivers**
  - GRID & XenServer Compatible Servers

- Shared GPU for Apps
  - XenApp 6.5 or XenApp 7.5
  - Bare Metal; vSphere; XenServer
  - NVIDIA graphics cards
  - Latest NVIDIA GRID vGPU **Drivers**
  - XenServer Compatible Servers

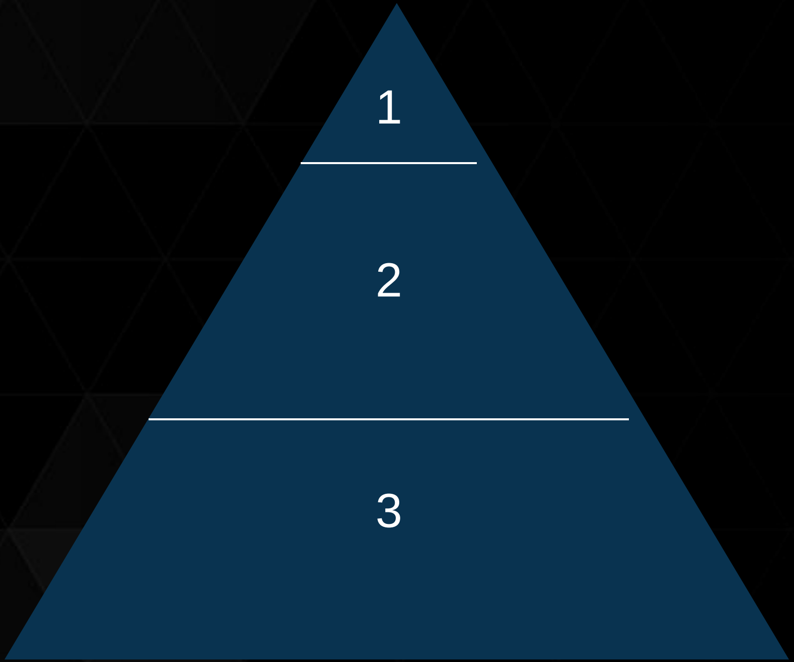
COMMON QUESTIONS

- XenDesktop VDI or RDS (XenApp)?
- Which NVIDIA card?
- If XenApp, bare metal or hypervisor?
- What server hardware?
- How many VMs per host? How many users per GPU?

BEFORE YOU BEGIN… ASK QUALIFYING QUESTIONS

1. Understand the target users
2. Segment the user population
3. Choose between VDI and RDS workloads
4. Choose the appropriate graphics card
5. Choose the server
6. Understand performance requirements & considerations

UNDERSTAND AND SEGMENT THE USER POPULATION



Tier 1: Professional users

(e.g. design engineers, radiologists)

- Top rendering performance
- 3D mouse support

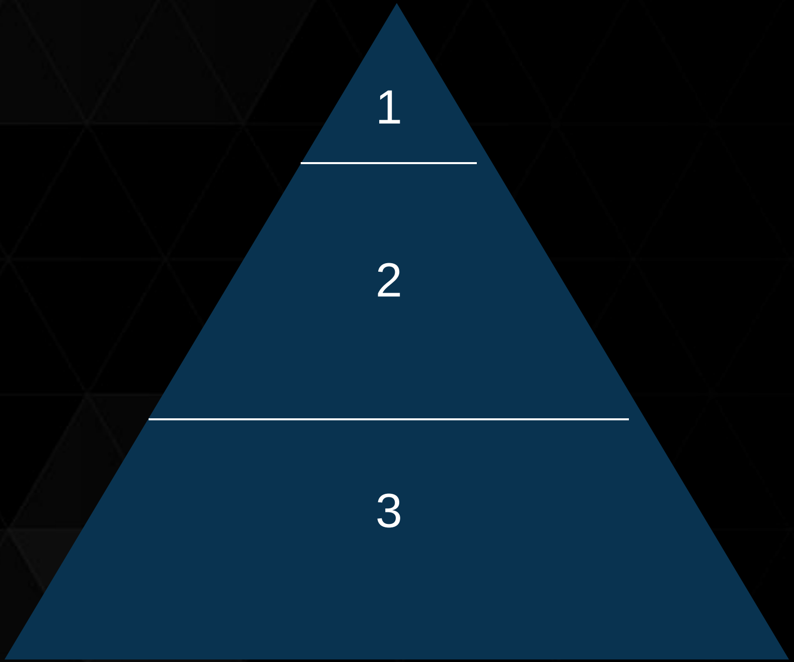
Tier 2: Power users

- Viewing of large 3D models, basic editing

Tier 3: Knowledge workers

- Limited use of 3D graphics today
- 2D apps, Aero effects of Windows, HD videos, PowerPoint slide transitions, etc.

USER SEGMENT DETERMINES BASIC APPROACH



Tier 1: Professional users

(e.g. design engineers, radiologists)

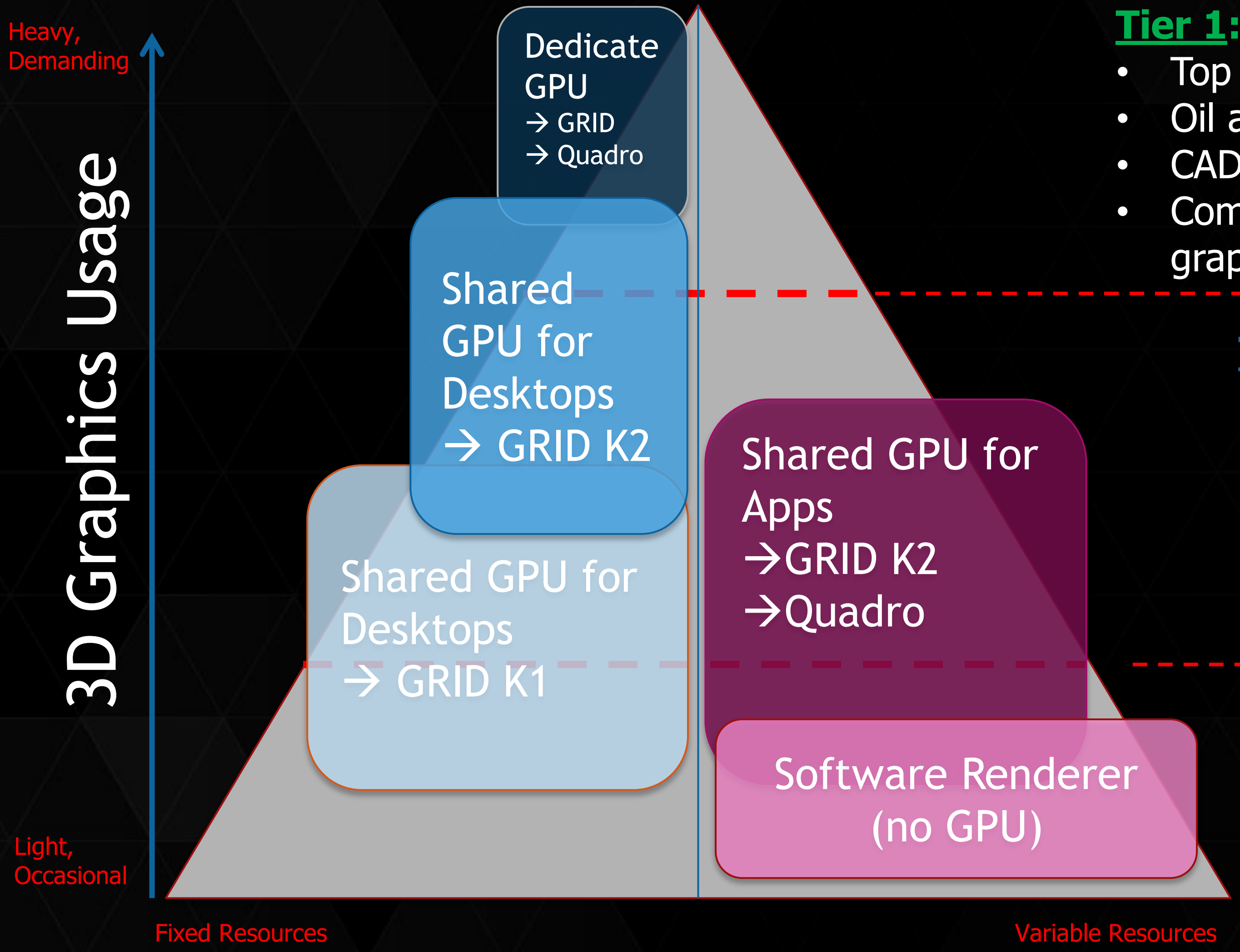
- **VDI workload** for best user experience
- Dedicated GPU or high-end vGPU profile

Tier 2: Power users

- **GPU sharing** for reasonable cost per user
- Choice of **VDI or RDS** workloads

Tier 3: Knowledge workers

- Software rasterizer or highly shared GPU



Tier 1: Design engineers

- Top rendering performance
- Oil and Gas, Aviation
- CAD/CAM, GIS
- Complex, mission-critical graphics

Tier 2: Power users

- Editor, Reviewer
- Compute Intensive Apps
- Mixed graphical and non-graphical apps

Tier 3: Knowledge workers

- Modern UI, MS Office; light 2D/3D
- No rendering

TIER 2 USERS: VDI OR RDS?

- Both approaches support:
  - GPU sharing with direct access to the graphics driver and hardware (no API intercept)
  - DirectX and OpenGL graphics acceleration
  - Adaptive H.264-based Deep Compression or pixel-perfect Lossless Compression
  - Delivery of full virtual desktop or seamless apps to multiple monitors

- Differences:

VDI - *Performance & Compatibility*
3D mouse support
Broadest app compatibility
CUDA and OpenCL support on bare metal but not yet supported by GRID vGPU

RDS - *Lowest Cost Per User*
Lowest cost (e.g. Microsoft licenses)
Apps must be RDS compatible
CUDA and OpenCL support is currently “experimental” pending field validation

RDS-COMPATIBLE PROFESSIONAL GRAPHICS APPS

- Some *examples* from autodeskandcitrix.com, Citrix Ready site, etc.
  (Note: AppDNA makes it easy to check XenApp compatibility)

- Lots of Autodesk apps, including:
  - AutoCAD
  - Inventor
  - Revit
  - Navisworks
- Bentley MicroStation
- Ansys Workbench and Fluent

- Dassault CATIA and 3D VIA Composer Player
- ESRI ArcGIS
- Intergraph SmartPlant 3D
- Adobe Photoshop (Creative Suite)
- SAP Right Hemisphere 3D
- Siemens Solid Edge and Teamcenter

	NVIDIA GRID K1	NVIDIA GRID K2
GPU	4 Kepler GPUs	2 High End Kepler GPUs
CUDA cores	768 (192 per GPU)	3072 (1,536 per GPU)
Memory Size	16GB DDR3 (4GB per GPU)	8GB GDDR5 (4GB per GPU)
OpenGL	up to 4.3	up to 4.3
DirectX	up to 11	up to 11
GRID vGPU support	XenServer 6.2 SP1	XenServer 6.2 SP1
User Density	up to 32 (64-96 per server)	up to 16 (32-48 per server)

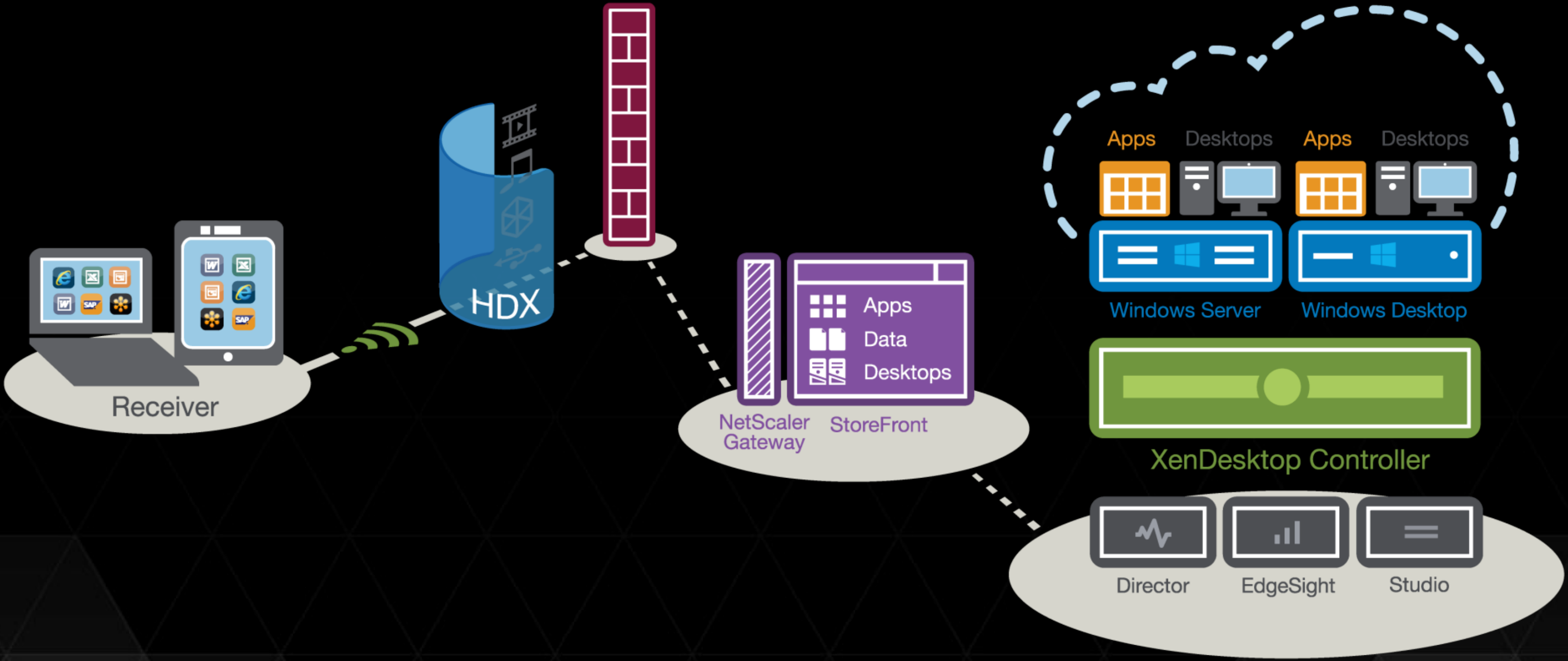
[1] *Number of users depends on software solution, workload, and screen resolution*

CHOOSING THE SERVER HARDWARE

CISCO	DELL	FUJITSU	hp	IBM
Cisco UCS C240 M3	Dell PowerEdge R720	Fujitsu Celsius C620	HP ProLiant WS460c Gen8	IBM iDataPlex dx360 M4
		Fujitsu Celsius R930	HP ProLiant SL250s Gen8	IBM Flex System

ASUS inspur Sugon SUPERMICRO TYAN

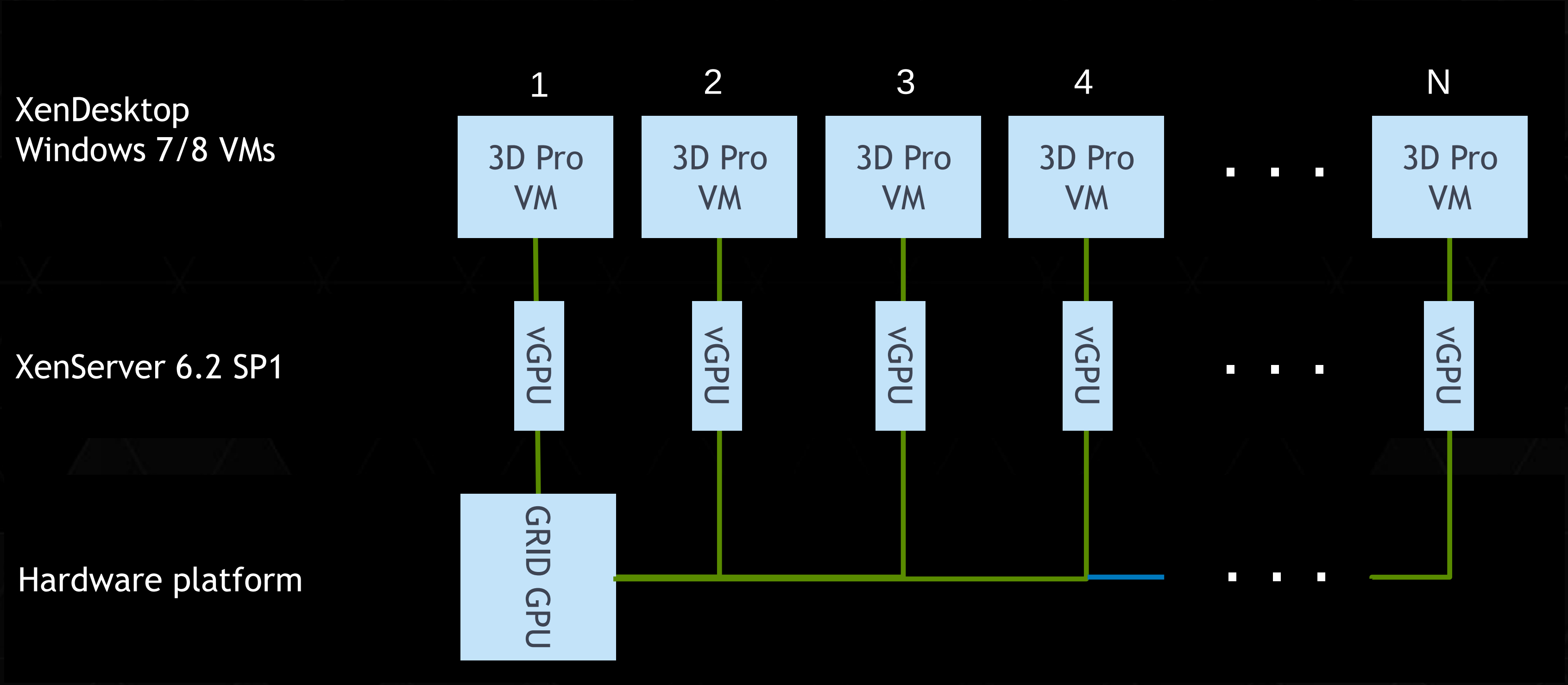
- Obviously, choose a server that supports your graphics card selection
- Check hypervisor compatibility!
  - XenServer: http://hcl.xensource.com/GPUPass-throughDeviceList.aspx
- Check eDocs for HDX 3D Pro minimum server requirements



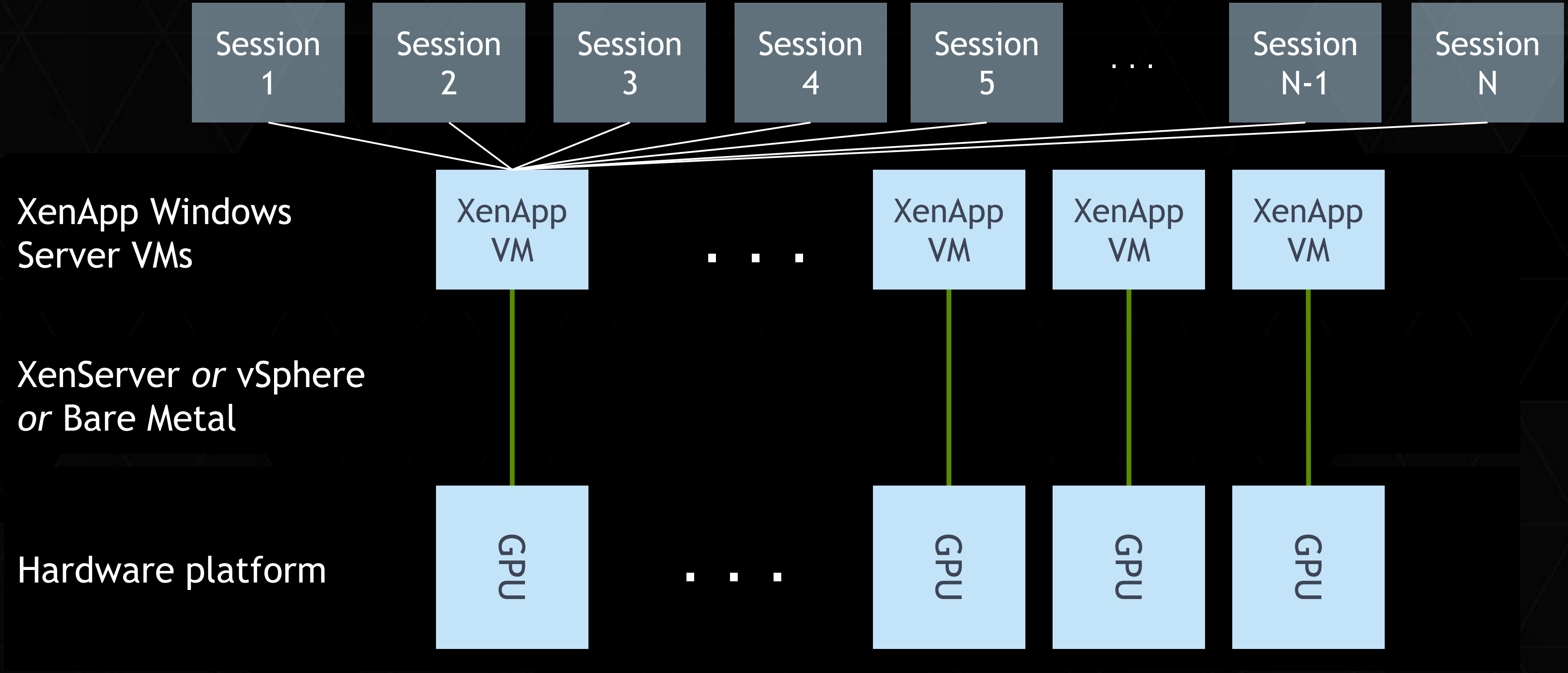
XENDESKTOP ARCHITECTURE

WINDOWS APPS AND DESKTOPS AS MOBILE SERVICES

GPU SHARING WITH XENDESKTOP



GPU SHARING WITH XENAPP



PERFORMANCE REQUIREMENTS VARY EVEN WITH THE SAME APPLICATION ..

CPU
Utilization

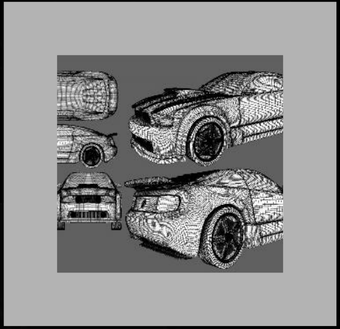
Memory
Utilization

Storage
Capacity & I/O

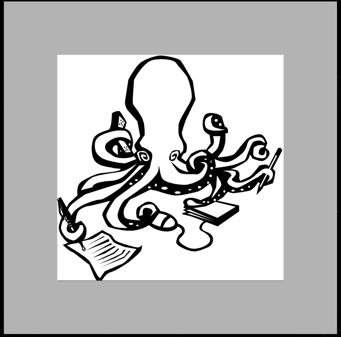
Network

GPU Core
Utilization

GPU Memory
Utilization



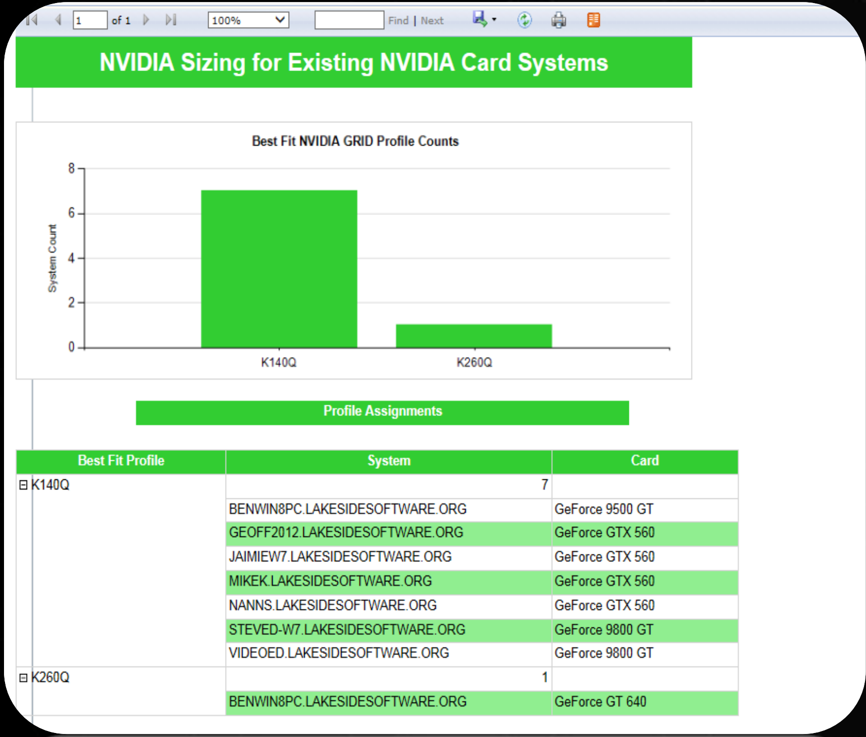
Application
Architecture



Usage
Concurrency

TOOLS NEEDED

- Citrix Director + Edgesight
- Citrix HDX Monitor (CTX135817)
- GPU-Z
- Task Manager
- Perfmon
  - CPU
  - Memory
  - Disk
  - Network
- Lakeside Software SysTrack



NVIDIA Sizing for Existing NVIDIA Card Systems

Profile Assignments

Best Fit Profile	System	Card
⊟ K140Q	7	
	BENWIN8PC.LAKESIDESOFTWARE.ORG	GeForce 9500 GT
	GEOFF2012.LAKESIDESOFTWARE.ORG	GeForce GTX 560
	JAIMIEW7.LAKESIDESOFTWARE.ORG	GeForce GTX 560
	MIKEK.LAKESIDESOFTWARE.ORG	GeForce GTX 560
	NANNS.LAKESIDESOFTWARE.ORG	GeForce GTX 560
	STEVED-W7.LAKESIDESOFTWARE.ORG	GeForce 9800 GT
	VIDEOED.LAKESIDESOFTWARE.ORG	GeForce 9800 GT
⊟ K260Q	1	
	BENWIN8PC.LAKESIDESOFTWARE.ORG	GeForce GT 640

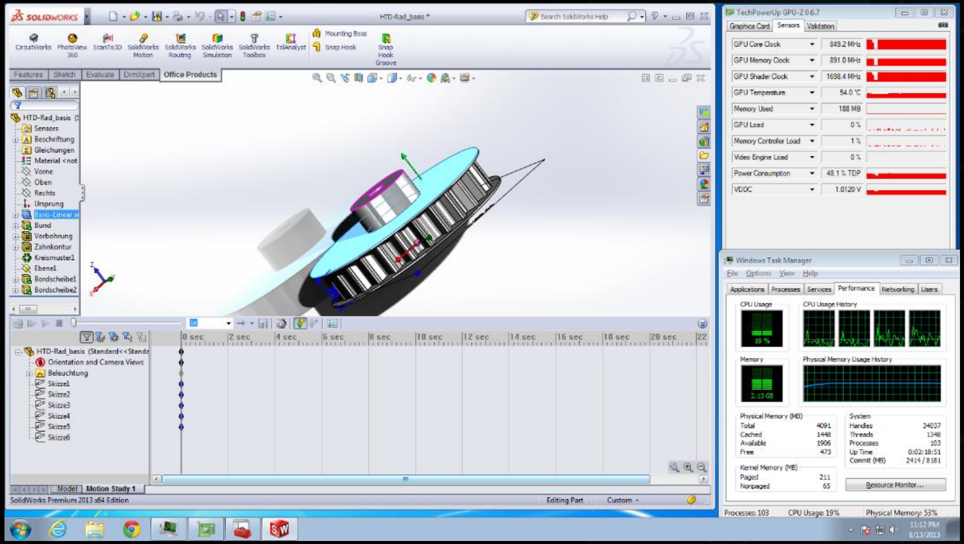
SAMPLE APP: DASSAULT SOLIDWORKS

4-vCPU Windows 7 VM

GPU Passthrough

NVIDIA K1 (192 cores GPU)

- **Performance Profile (Average)**
- CPU Load: 18% (41% peak)
- GPU Load: 5% (25% peak)
- GPU Memory: 188 MB (net 64MB)
- Network Out: 752 Kbps (2.3 Mbps peak)
- Network In: 43 Kbps
- Disk Reads/Sec: <1
- Disk Writes/Sec: 4 (21 peak)



GPU Core Clock	849.2 MHz
GPU Memory Clock	891.0 MHz
GPU Shader Clock	1698.4 MHz
GPU Temperature	54.0 °C
Memory Used	188 MB
GPU Load	0 %
Memory Controller Load	1 %
Video Engine Load	0 %
Power Consumption	48.1 % TDP
VDDC	1.0120 V

Physical Memory (MB)	
Total	4091
Cached	1448
Available	1906
Free	473

Kernel Memory (MB)	
Paged	211
Nonpaged	65

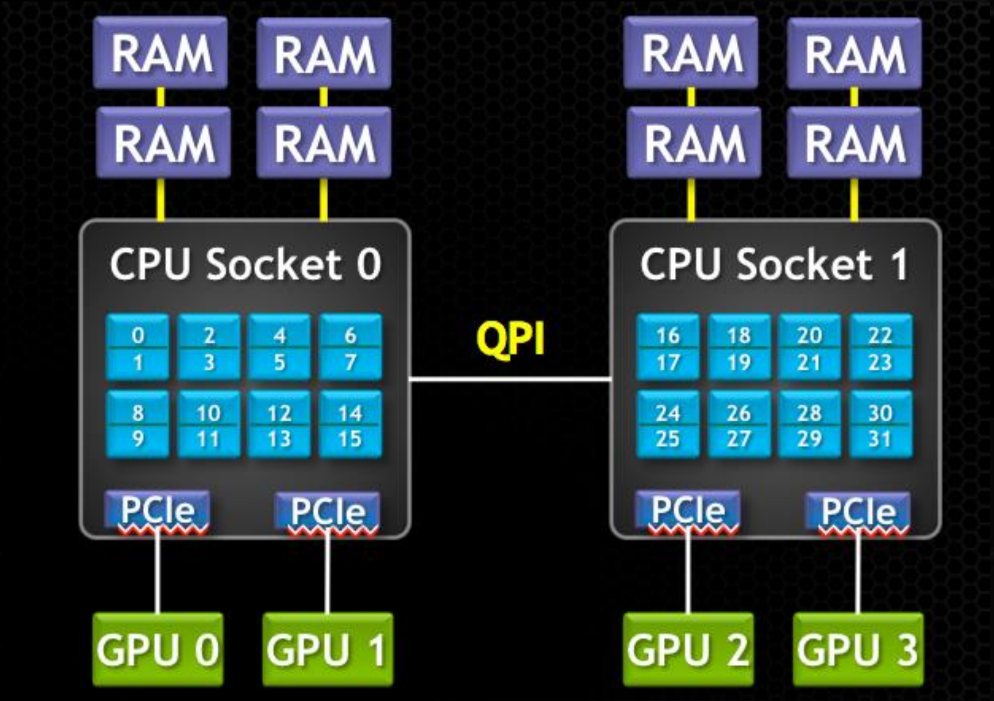
System	
Handles	34037
Threads	1348
Processes	103
Up Time	0:02:18:51
Commit (MB)	2414 / 8181

Processes: 103   CPU Usage: 19%   Physical Memory: 53%

NUMA AFFINITY IMPROVES PERFORMANCE

Avoid the overhead of going through the CPU interconnect

- Improves performance by up to 15% depending on the application and use case



Courtesy of NVIDIA Corp.

STORAGE CONSIDERATIONS

- IOPS
  - Initial data load – MBs to GBs of data, 10s to 100s IOPS
  - Steady state – 10-200 IOPS (SolidWorks, AutoDesk Inventor, AutoDesk Revit, Right Hemisphere, GoogleEarth)
- Storage
  - Local SSD
  - SAN or NAS

NETWORK CONSIDERATIONS – BANDWIDTH CONSUMPTION

- Bandwidth requirement is use-case specific
- Range is between average of 300 Kbps to 2 Mbps
- Case examples
  - Custom imaging application - 300-500 Kbps
  - GoogleEarth - ~1-1.5 Mbps
  - Siemens NX for electronics use case - ~1Mbps

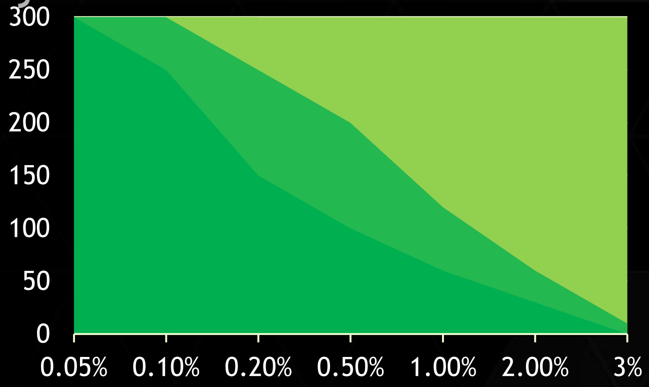
NETWORK CONSIDERATIONS – NETWORK LATENCY

Latency (ms)	Rating	Comments
<= 50	Best	Suitable for most demanding use cases; for example: animation or panning complex maps
51 to 100	Better	Suitable up to power user requirements; non-interactive and "viewing" workflows
101 to 150	Good	Provides very usable sessions but some use cases may find it sluggish.
151 to 300	Acceptable	Remote sites like India and China may find it acceptable.
>301	Use-case specific	Some use cases may still find it usable such as technical reviewers or writers working from remote locations.

LOOKING AHEAD… FRAMEHAWK INNOVATIONS

Enhancements to the HDX stack will benefit 3D graphics users on difficult network connections

- Human heuristic driven graphics display
- Image/pattern recognition
- Instantly interruptible graphics layer
- QoS signals amplifier
- Time-based heat map



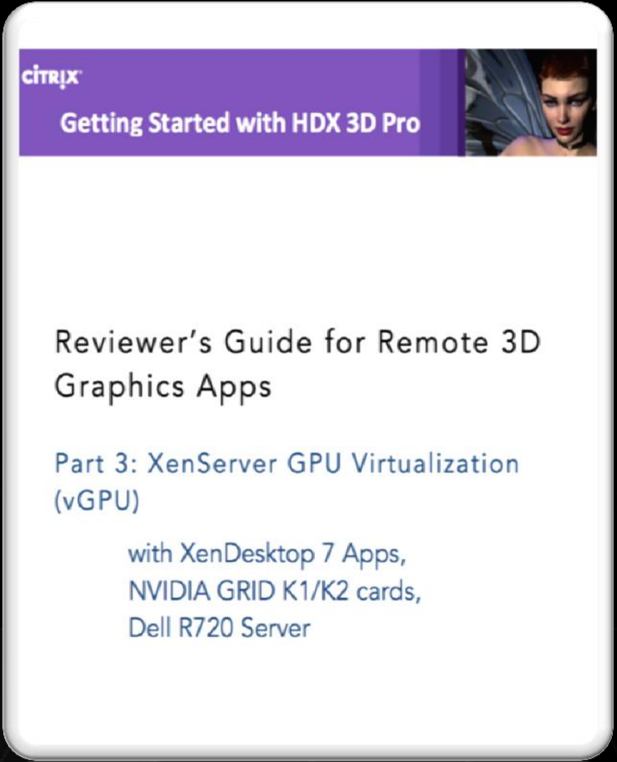
Framehawk will extend HDX to support even more demanding network conditions

All tests are conducted at 250ms
and mobile scenarios varying from 5% to 50% loss

RECAP

1. Understand the target users
2. Segment the user population
3. Choose between VDI and RDS workloads
4. Choose the appropriate graphics card
5. Choose the server
6. Understand the performance requirements & considerations

ADDITIONAL INFO ABOUT CITRIX HDX 3D PRO



CITRIX

Getting Started with HDX 3D Pro

Reviewer's Guide for Remote 3D Graphics Apps

Part 3: XenServer GPU Virtualization (vGPU)

with XenDesktop 7 Apps,
NVIDIA GRID K1/K2 cards,
Dell R720 Server

http://blogs.citrix.com/2013/09/10/new-reviewers-guide-for-xendesktop-7-hdx-3d-pro-graphics-on-both-xenserver-and-vsphere/

http://blogs.citrix.com/2013/12/24/scripting-automating-the-testing-of-graphic-intensive-gpu-workloads/

Release Notes and Admin Guide, on http://www.citrix.com/go/vGPU

Design Guide for Virtual Design Engineering

Work better. Live better.