GPU TECHNOLOGY CONFERENCE

PLANNING FOR DENSITY AND PERFORMANCE IN VDI WITH NVIDIA GRID

JASON SOUTHERN

SENIOR SOLUTIONS ARCHITECT FOR NVIDIA GRID

AGENDA

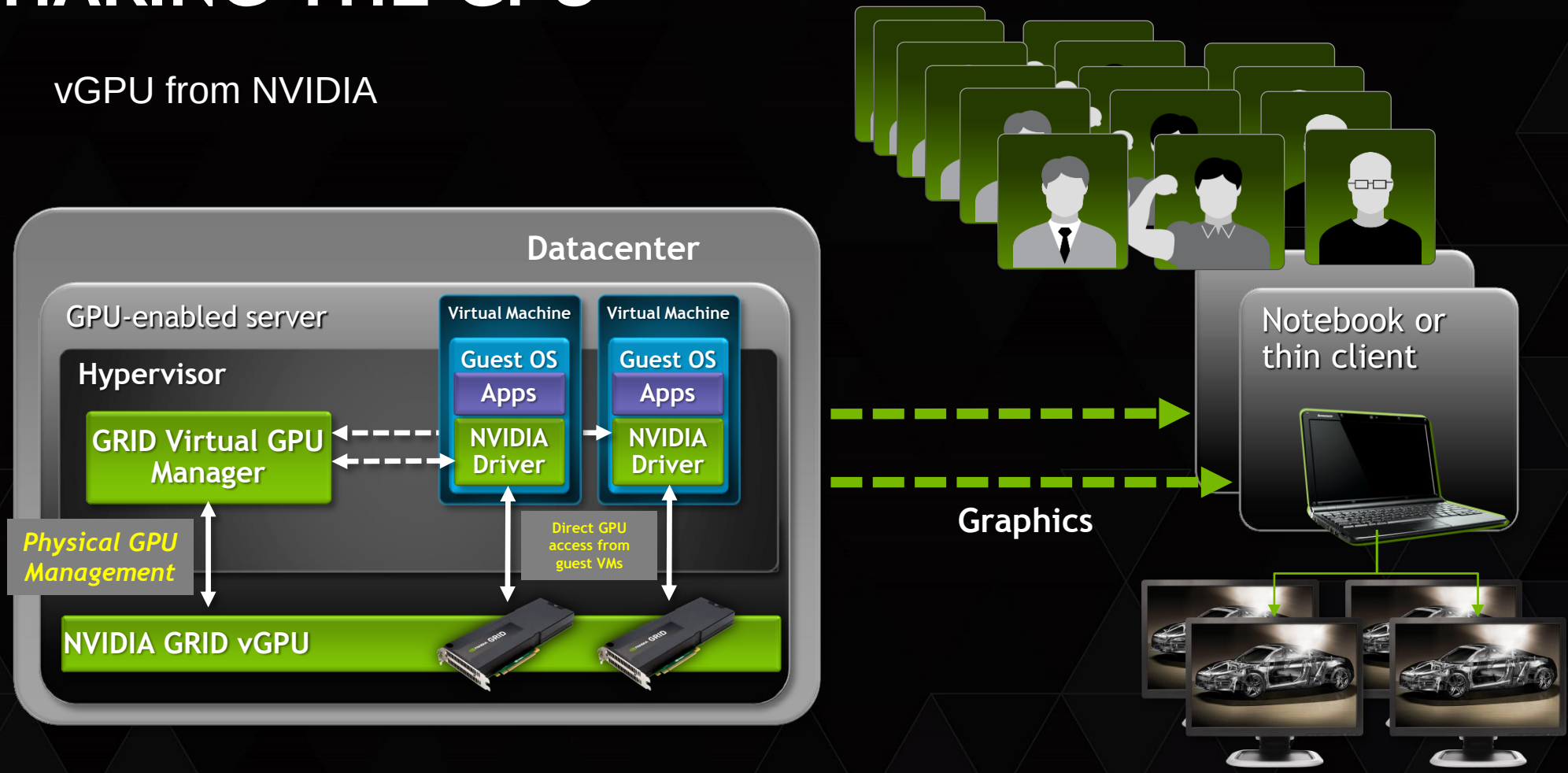
- Recap on how vGPU works
- Planning for Performance
  - Design considerations
  - Benchmarking
- Optimizing for Density

Nvidia vGPU

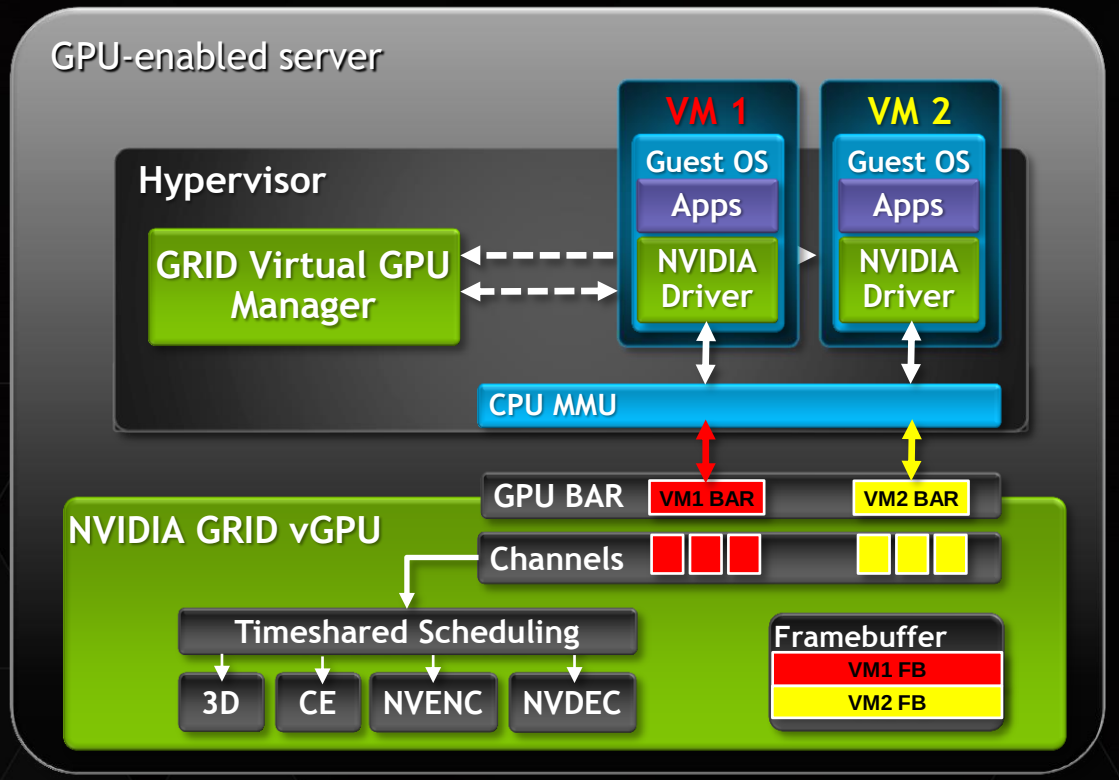
- recap

SHARING THE GPU

vGPU from NVIDIA



VIRTUAL GPU RESOURCE SHARING



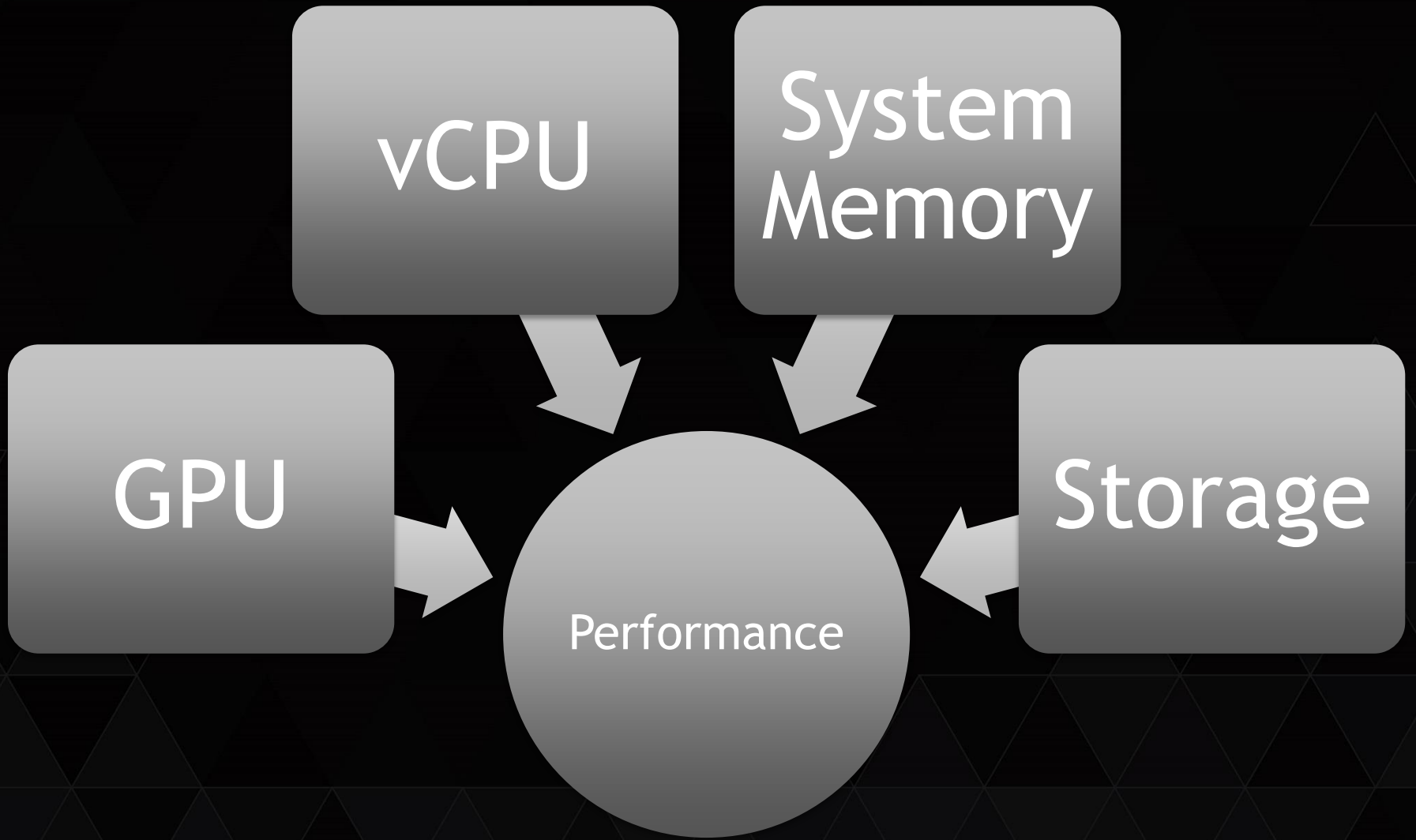
- Frame buffer
  - Fixed allocation
  - Allocated at VM startup
- GPU Engines

  Timeshared among VMs, like multiple contexts on single OS

  Dedicated secure data channels between VM & GPU

Building for Performance

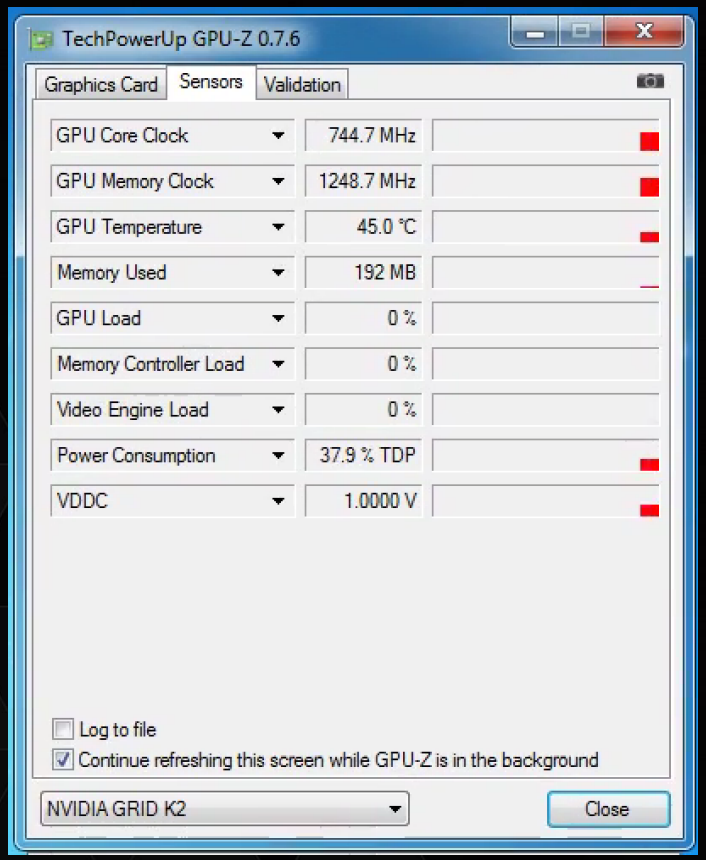
WHAT AFFECTS OVERALL PERFORMANCE

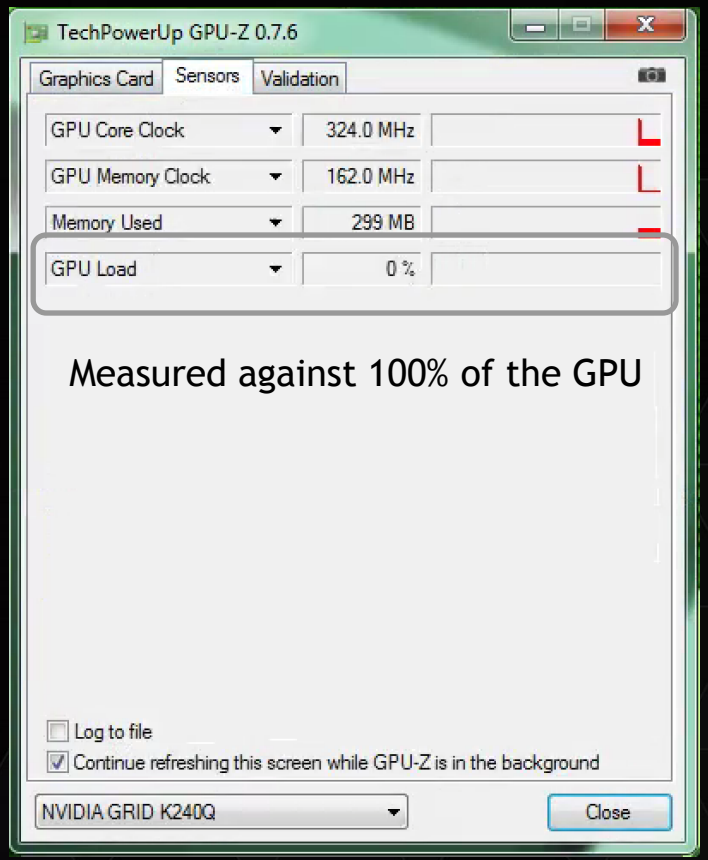


HOW DO WE CHECK GPU UTILIZATION?

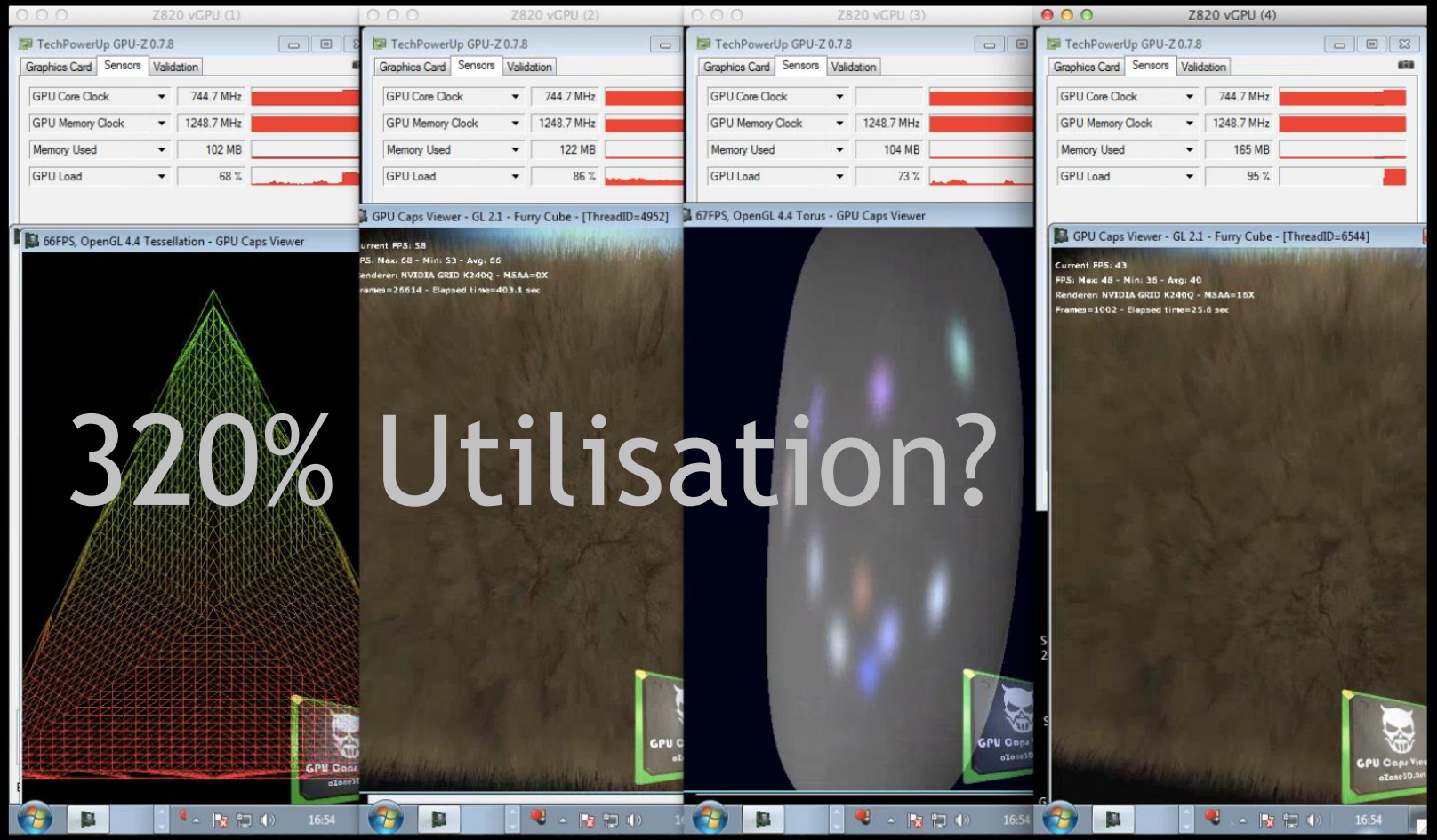
- Nvidia-SMI
  - CLI
  - Realtime & Looping
- Perfmon
  - GUI
  - Realtime & logging
- GPU-Z
  - GUI
  - Realtime & Log to File
- Process Explorer
  - Per process information on utilisation
- GPUShark
  - Basic GUI
  - Realtime
- Lakeside Systrack / LWL Stratusphere
  - Detailed historical reporting

MONITORING PASSTHROUGH VS VGPU





BE CAREFUL THOUGH…



GPU-Z

GPUcapsviewer

- VM1 – 68%
- VM2 – 86%
- VM3 – 73%
- VM4 – 95%

ASSESSMENT TOOLS

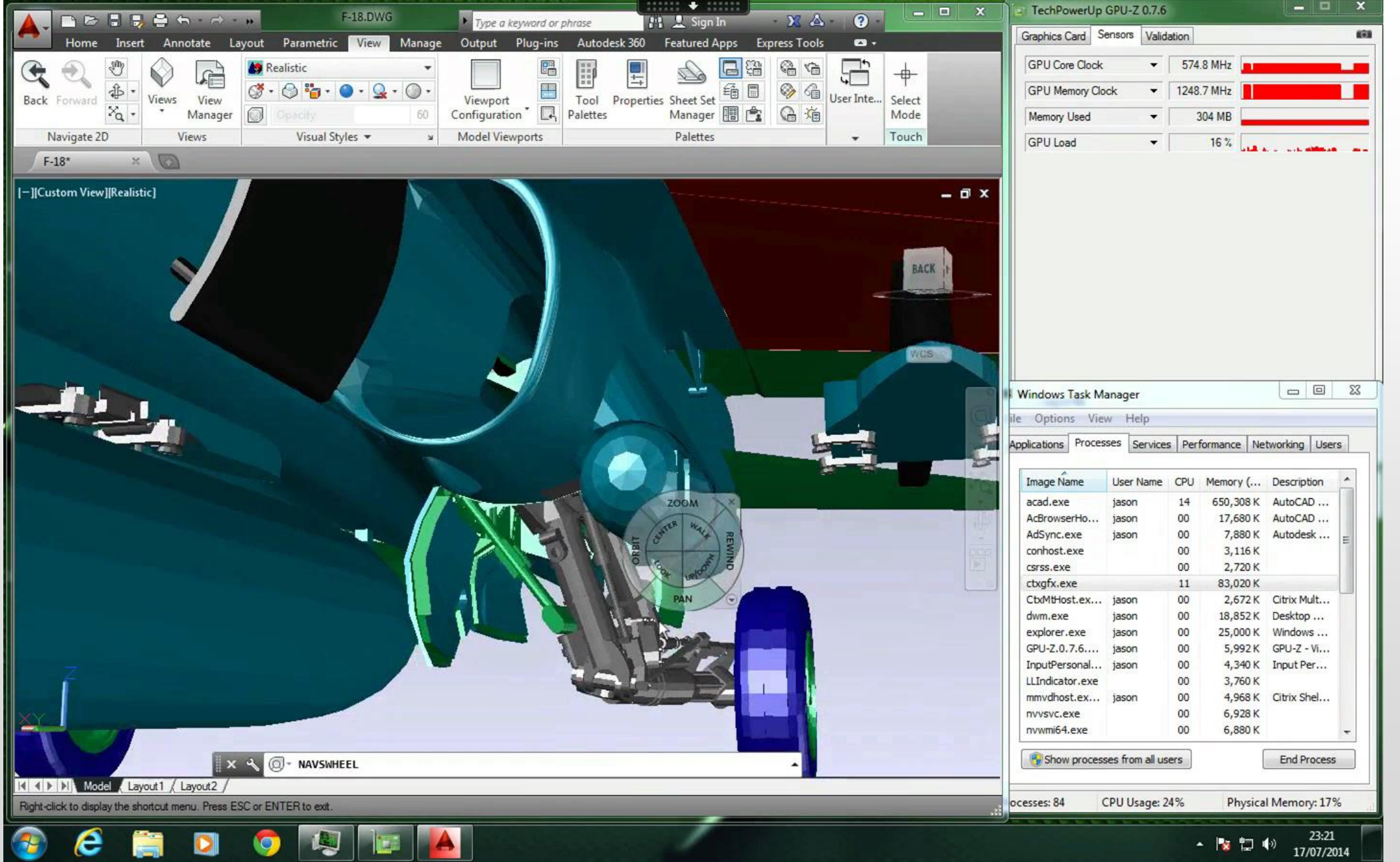
- Long term assessment data allows you to plan for the peak loads.
- GPU usage is often in bursts, plan for the peak not the mean.

- Use assessment tools that track GPU info e.g.

  - Lakeside Systrack 7
  - Liquidware Labs Stratusphere FIT

VCPU’S

- Allow at least one for the Encoder (HDX or PCoIP)
- Allow at least one for the OS
- The rest are for the application(s)
  - How many did the workstations have?
  - How demanding is the application itself?


F-18.DWG
Type a keyword or phrase
Sign In
Home Insert Annotate Layout Parametric View Manage Output Plug-ins Autodesk 360 Featured Apps Express Tools
Back Forward
Views
View Manager
Realistic
Viewport Configuration
Tool Palettes
Properties
Sheet Set Manager
User Inte...
Select Mode
Navigate 2D
Views
Visual Styles
Model Viewports
Palettes
Touch
F-18*
[−][Custom View][Realistic]
BACK
WCS
ZOOM
CENTER
WALK
REWIND
ORBIT
LOOK
UP/DOWN
PAN
NAVSWHEEL
Model
Layout1
Layout2
Right-click to display the shortcut menu. Press ESC or ENTER to exit.
TechPowerUp GPU-Z 0.7.6
Graphics Card
Sensors
Validation
GPU Core Clock 574.8 MHz
GPU Memory Clock 1248.7 MHz
Memory Used 304 MB
GPU Load 16 %
Windows Task Manager
File Options View Help
Applications
Processes
Services
Performance
Networking
Users
Image Name
User Name
CPU
Memory (...
Description
acad.exe jason 14 650,308 K AutoCAD ...
AcBrowserHo... jason 00 17,680 K AutoCAD ...
AdSync.exe jason 00 7,880 K Autodesk ...
conhost.exe 00 3,116 K
csrss.exe 00 2,720 K
ctxgfx.exe 11 83,020 K
CtxMtHost.ex... jason 00 2,672 K Citrix Mult...
dwm.exe jason 00 18,852 K Desktop ...
explorer.exe jason 00 25,000 K Windows ...
GPU-Z.0.7.6.... jason 00 5,992 K GPU-Z - Vi...
InputPersonal... jason 00 4,340 K Input Per...
LLIndicator.exe 00 3,760 K
mmvdhost.ex... jason 00 4,968 K Citrix Shel...
nvvsvc.exe 00 6,928 K
nvwmi64.exe 00 6,880 K
Show processes from all users
End Process
ocesses: 84
CPU Usage: 24%
Physical Memory: 17%
23:21
17/07/2014


MARTINEZ
11
RC
Orange

SYSTEM MEMORY

- => GPU Memory
  2GB of System RAM & 4GB GPU Memory = Bottleneck!

- Memory overcommit / ballooning etc is not recommended.

PASSTHROUGH OR VGPU

When do I really need to use Passthrough?

- CUDA
- Computational Usage – GPGPU
- PhysX

- Troubleshooting vGPU issues
- Driver simplification
  - Kx80Q

CUDA – WHAT IS IT

- NVIDIA's parallel computing architecture that enables dramatic increases in computing performance by harnessing the power of the GPU

- Applications & their features that use CUDA

  http://www.nvidia.com/object/gpu-accelerated-applications.html

Benchmarking

BENCHMARKING

- Remember – you're benchmarking the entire VM, not just the GPU
- All of these have an impact on the result.

  - GPU
  - CPU
  - RAM
  - DISK

- Don't overlook User Experience testing.
  - Benchmarks are just numbers, user acceptance is king.

BENCHMARKING TOOLS

- CADalyst
  - For AutoCAD workloads
  http://www.cadalyst.com/benchmark-test

- 3D Mark 11
  - Generic DirectX benchmarking
  http://www.futuremark.com/benchmarks/3dmark11

- SPECViewperf 11
  - OPENGL benchmarking tool
  - Has industry & application specific modules available
  - Version 12 has issues with virtualisation at present..
  http://www.spec.org/gwpg/gpc.static/vp11info.html

Frame Rate Limiter & VSYNC

FRAME RATE LIMITER

- For vGPU we implement a frame Rate Limiter (FRL)
- Used in vGPU to balance performance across multiple vGPUs executing on the same physical GPU.
- FRL imposes a max frames-per-second that vGPU will render at in a VM.
  - Q profiles render at 60fps max
  - non Q profiles are limited to 45fps max

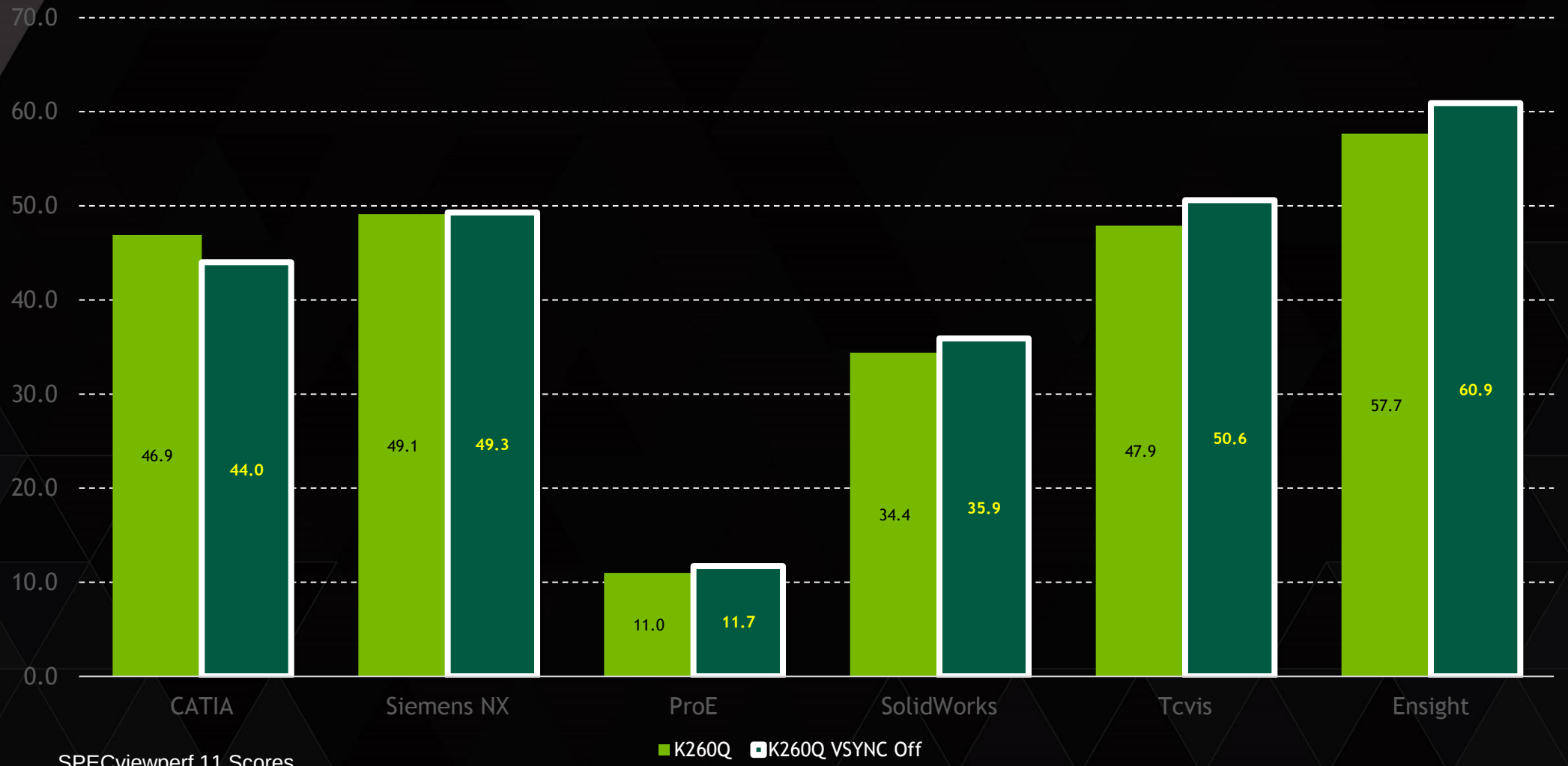
With FRL

Without FRL

VSYNC

- Setting is modified by applications or manually performed via the NVIDIA Control Panel
- Default setting allows the application to set the VSYNC policy
- Setting the VSYNC to “on” will synchronize the frame rate to 60Hz / 60 FPS for both pass-through and vGPU
- Setting the VSYNC to “off” will allow the GPU to render as many frames as possible
  - In vGPU profiles, this setting does not override the FRL

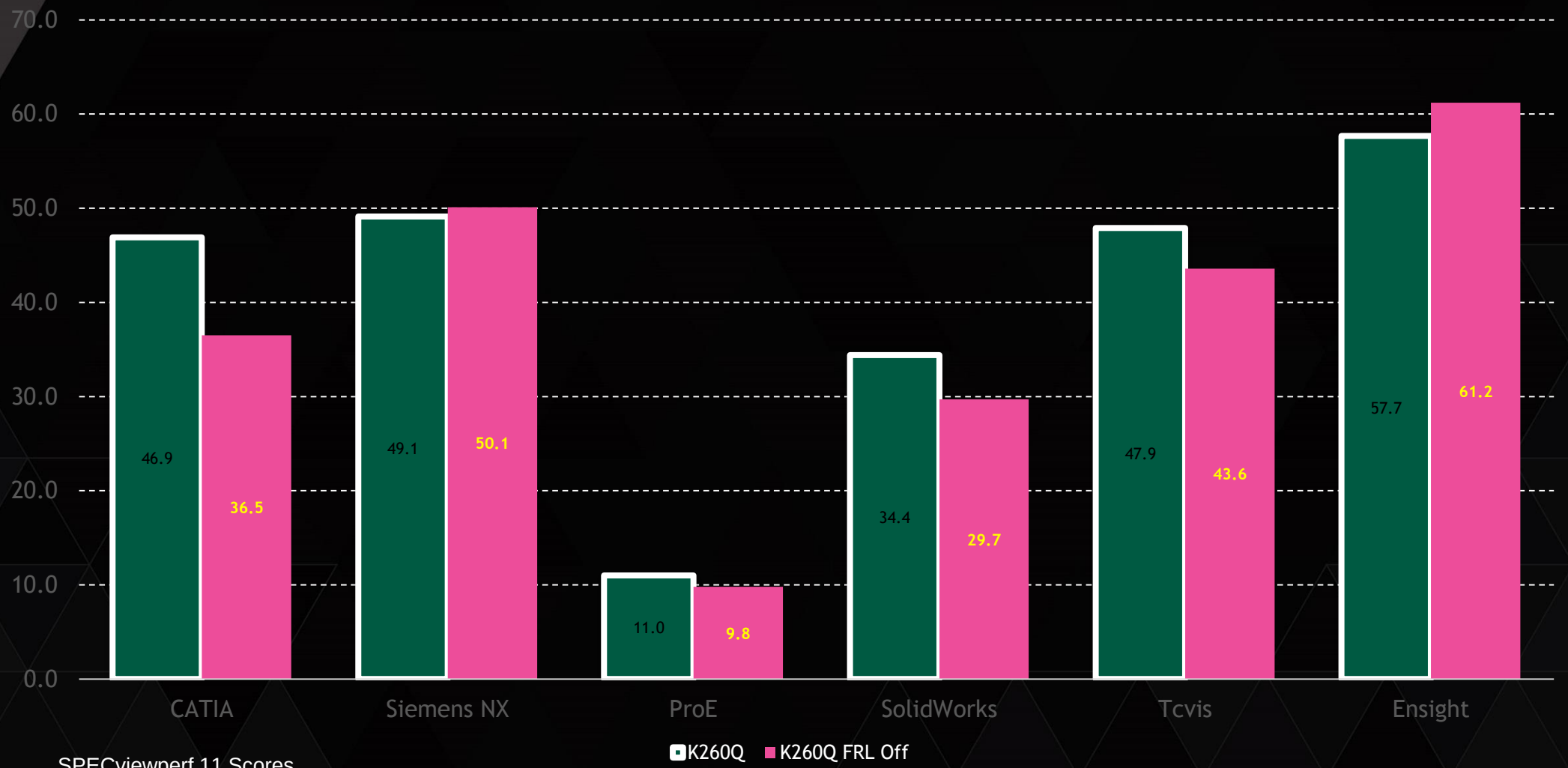
VSYNC EFFECT ON VGPU - SINGLE VM



SPECviewperf 11 Scores

GPU TECHNOLOGY CONFERENCE

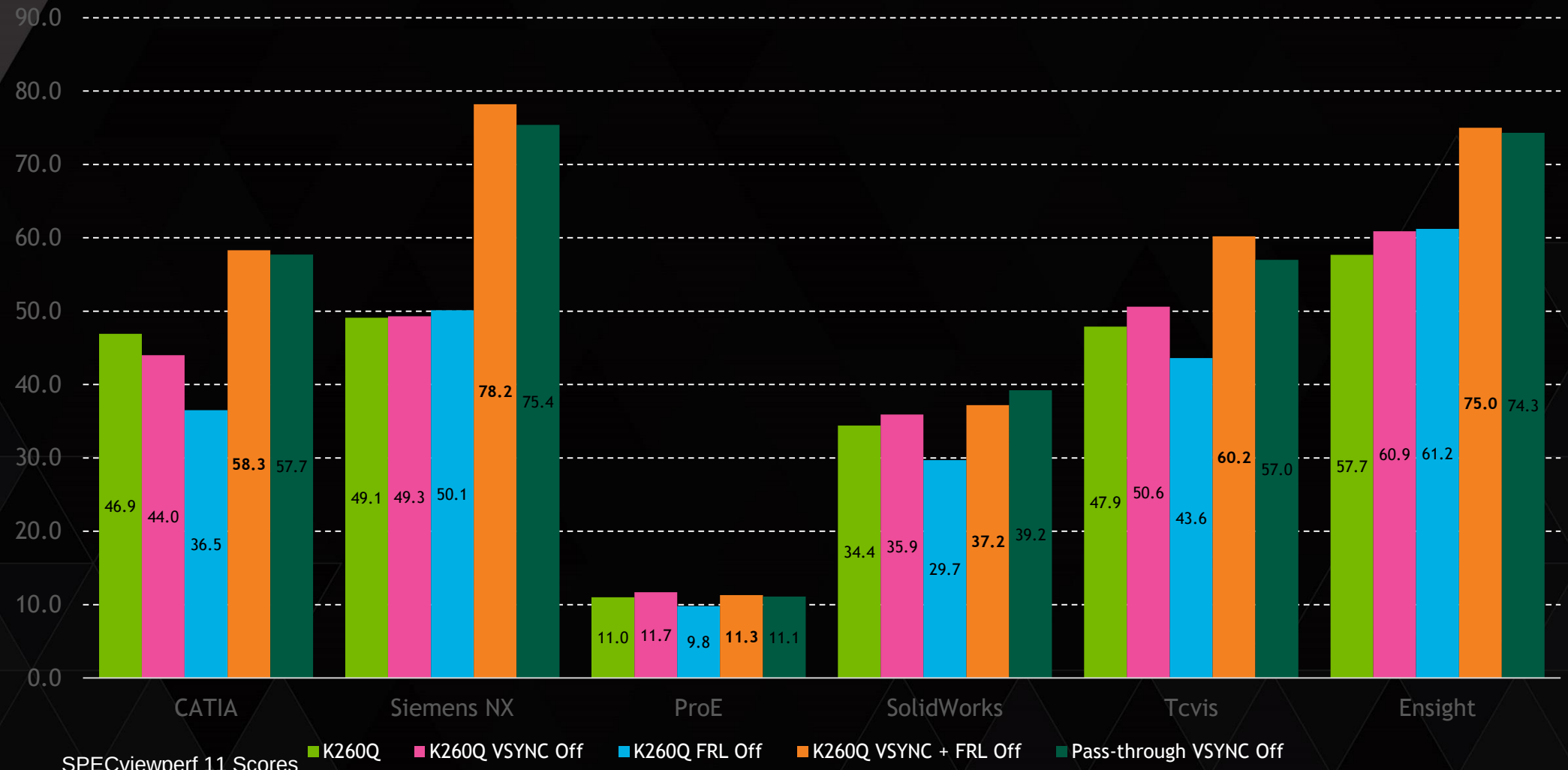
FRL EFFECT ON VGPU – SINGLE VM



SPECviewperf 11 Scores

GPU TECHNOLOGY CONFERENCE

VSYNC + FRL EFFECT ON VGPU



Optimizing for Density

- Am I using the right profile?

COMPARING QUADRO TO VGPU

Quadro K6000
2880 CUDA cores
12GB

Quadro K5000
1536 CUDA cores
4GB

Quadro K4000
768 CUDA cores
3GB

Quadro K2000
384 CUDA cores
2GB

Quadro K600
192 CUDA cores
1GB

Quadro 410
192 CUDA cores
512MB

Pass-through

GRID K2
2x 1536 CUDA cores
2x 4GB

GRID K1
4x 192 CUDA cores
4x 4GB

vGPU

GRID K260Q
2x 1536 CUDA cores
4x 2GB

GRID K240Q
2x 1536 CUDA cores
8x 1GB

GRID K140Q
4x 192 CUDA cores
16x 1GB

vGPU Profiles In Current Driver

Board	vGPU type	vGPUs per board	vGPUs per GPU	Per virtual GPU		
				FB	Heads	Max Res
GRID K1	GRID K120Q	32	8	512M	2	2560x1600
	GRID K140Q	16	4	1G	2	2560x1600
	GRID K160Q	8	2	2G	4	2560x1600
	GRID K180Q	4	1	4G	4	2560x1600

Board	vGPU type	vGPUs per board	vGPUs per GPU	Per virtual GPU		
				FB	Heads	Max Res
GRID K2	GRID K220Q	16	8	512M	2	2560x1600
	GRID K240Q	8	4	1G	2	2560x1600
	GRID K260Q	4	2	2G	4	2560x1600
	GRID K280Q	2	1	4G	4	2560x1600

What does the Q mean?

GRID K260Q
2GB framebuffer
4 heads, 2560x1600

ENGINEER
DESIGNER

GRID K240Q
1GB framebuffer
2 heads, 2560x1600

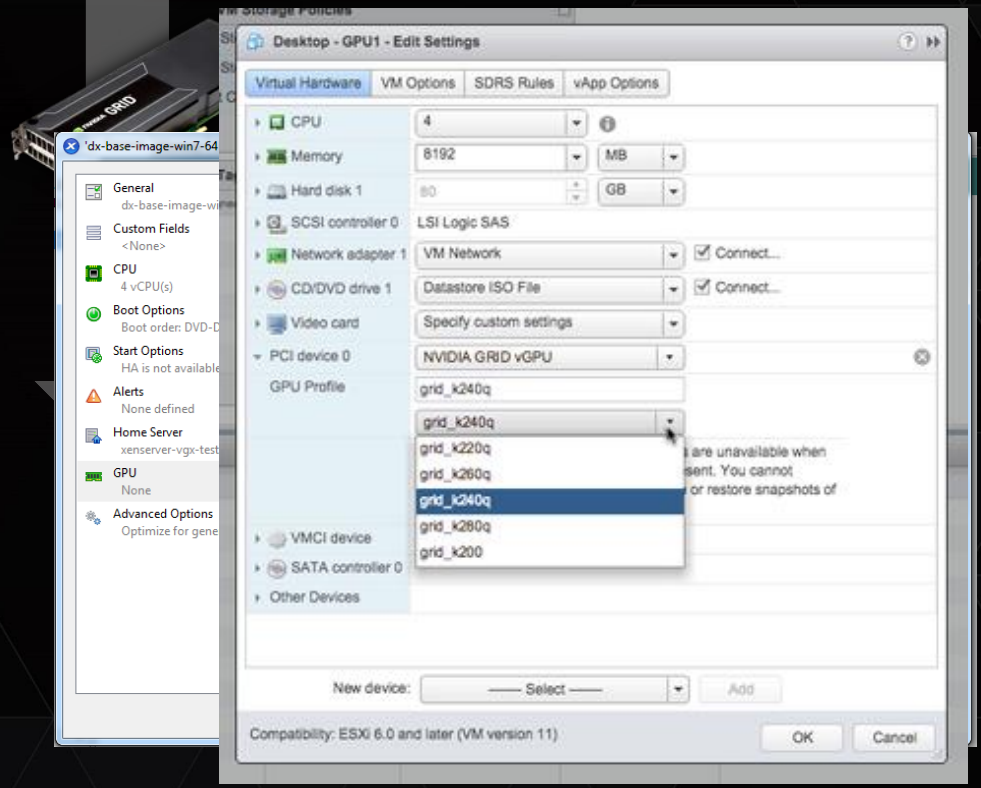
POWER USER

GRID K220Q
512MB framebuffer
2 heads, 1920x1200

KNOWLEDGE
WORKER

GRID K2

2 high-end Kepler GPUs
3072 CUDA cores (1536 / GPU)
8GB GDDR5 (4GB / GPU)



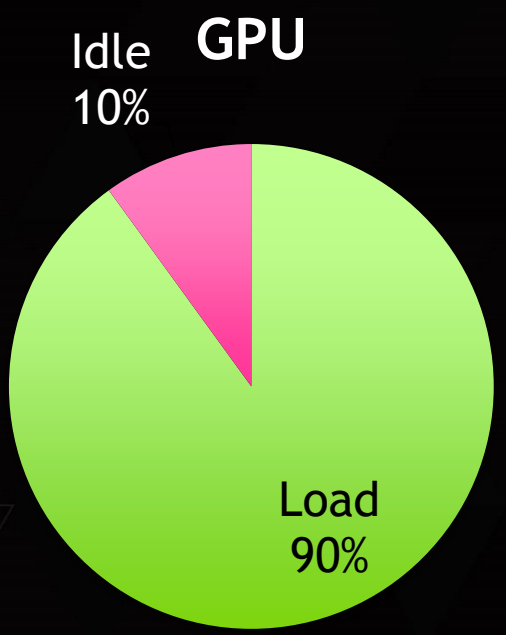
LET'S CONSIDER A SCENARIO.

- An organisation has trialled K1's in passthrough on dual displays
  - Performance is perfect, but they want better density from their server purchase if possible.

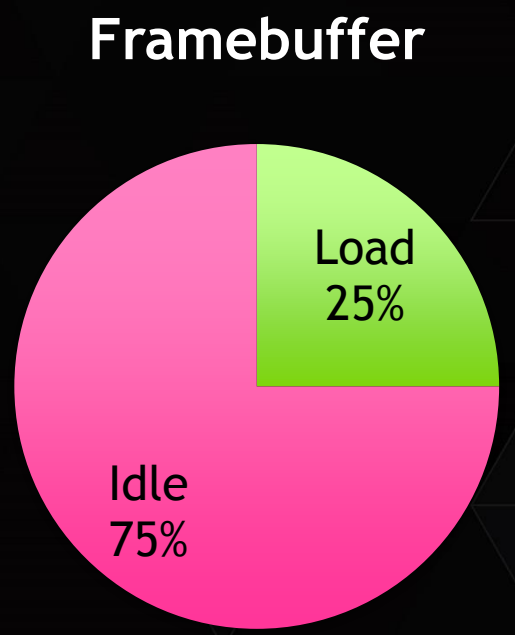
  -2 K1 cards in a chassis = 8 Users in pass-through.

- Is there a way to get more users on the server with the same or better performance?

IT DEPENDS ON THE PEAK UTILIZATION



90% of the GPU in use
vGPU on K1 not an option



1 GB Framebuffer in use
3 GB going to waste.

VGPU OPTIONS ON A K2 CARD.

Card	Physical GPUs	Virtual GPU	Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution	Maximum vGPUs	
							per GPU	per Board
GRID K2	2	GRID K260Q				2560x1600	2	4
GRID K2	2	GRID K240Q	Entry-Level Designer	1024	2	2560x1600	4	8
GRID K2	2	GRID K220Q		512	2	2560x1600	8	16

No Density improvement – 4 VM's per card

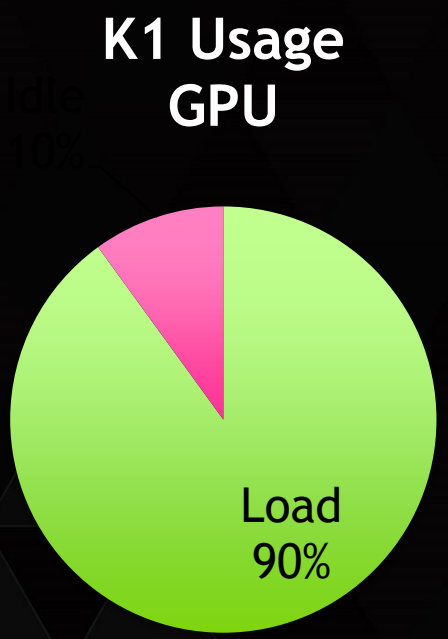
Sufficient Guaranteed GPU capacity but too little Framebuffer < 1Gb

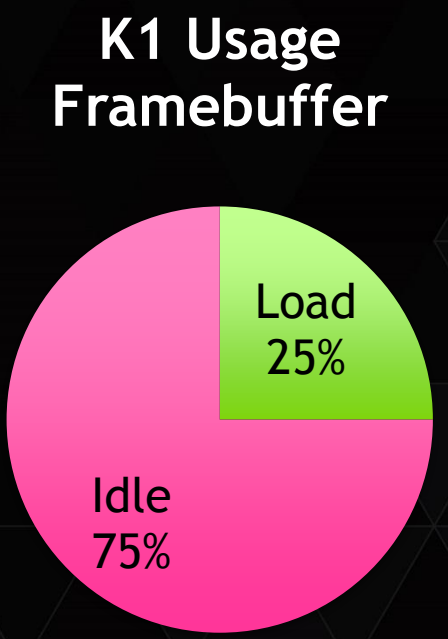
K1 – 192 Cores per GPU
K2 – 1536 Cores per GPU

So, let's assume that K220Q profiles have similar minimum GPU resources to K1 in pass-through

THE GOLDILOCKS PROFILE?

Card	Physical GPUs	Virtual GPU	Use Case	Frame Buffer (MB)	Virtual Display Heads	Maximum Resolution	Maximum vGPUs	
							per GPU	per Board
GRID K2	2	GRID K240Q	Entry-Level Designer	1024	2	2560x1600	4	8

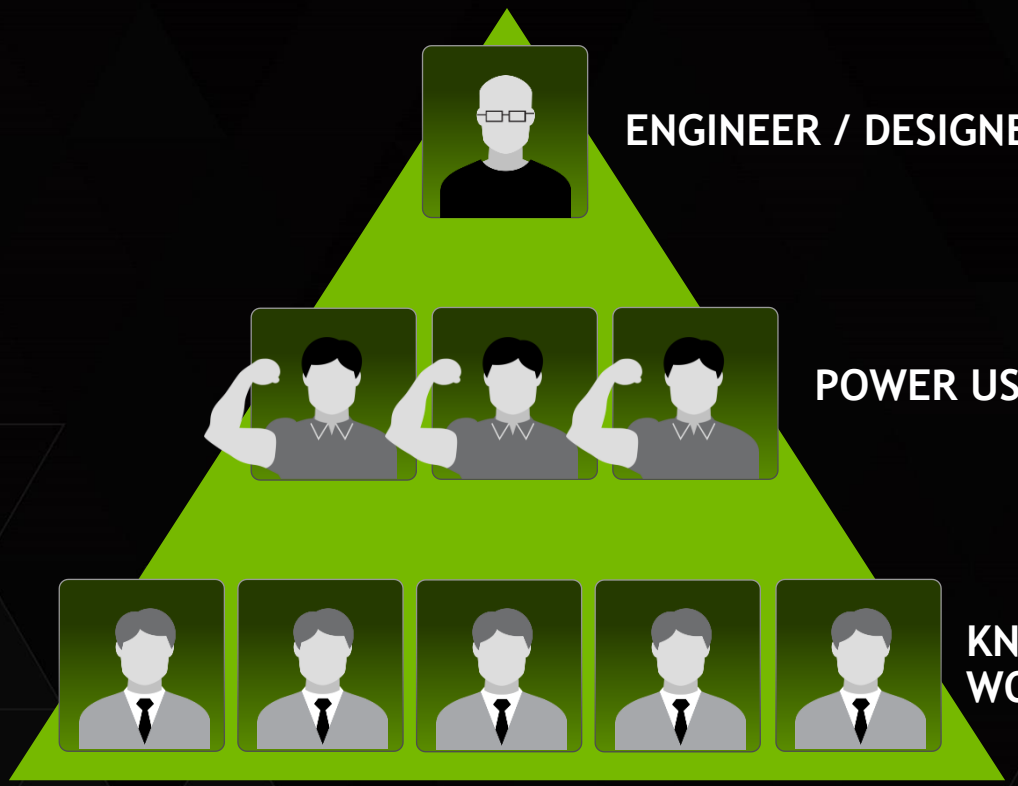




POTENTIAL SOLUTION

- K2 with 240Q profile would
  - Double the user density in the chassis to 16
  - Increased GPU performance
  - CAPEX reduction due to less chassis' needed.

Remember, this is just the start…



GRID K2

- High-end Kepler GPUs
- 3072 CUDA cores (1536 / GPU)
- 8GB GDDR5 (4GB / GPU)

GRID K1

- Entry Kepler GPUs
- 768 CUDA cores (192 / GPU)
- 16GB DDR3 (4GB / GPU)

One Last thing…

- Impact of Remoting Protocols


Recycle Bin
Benchmarks
Simulators
Fraps
REDTurbineDemo
Real-time viewport rendering
No user interaction in benchmark mode, otherwise press 'h' for help
2
ms
31%
315MB
47C
50.037W
12:11
18/07/2014

THANK YOU

JOIN THE CONVERSATION

#GTC15