GPU TECHNOLOGY CONFERENCE

April 4-7, 2016 | Silicon Valley

BENCHMARKING GRAPHICS INTENSIVE APPLICATION ON VMWARE HORIZON VIEW USING NVIDIA GRID VGPU

Manvender Rawat, NVIDIA

Lan Vu, VMware

PRESENTED BY

NVIDIA.

AGENDA

Overview of VMware Horizon 7 and NVIDIA GRID 2.0

How to Size VMs

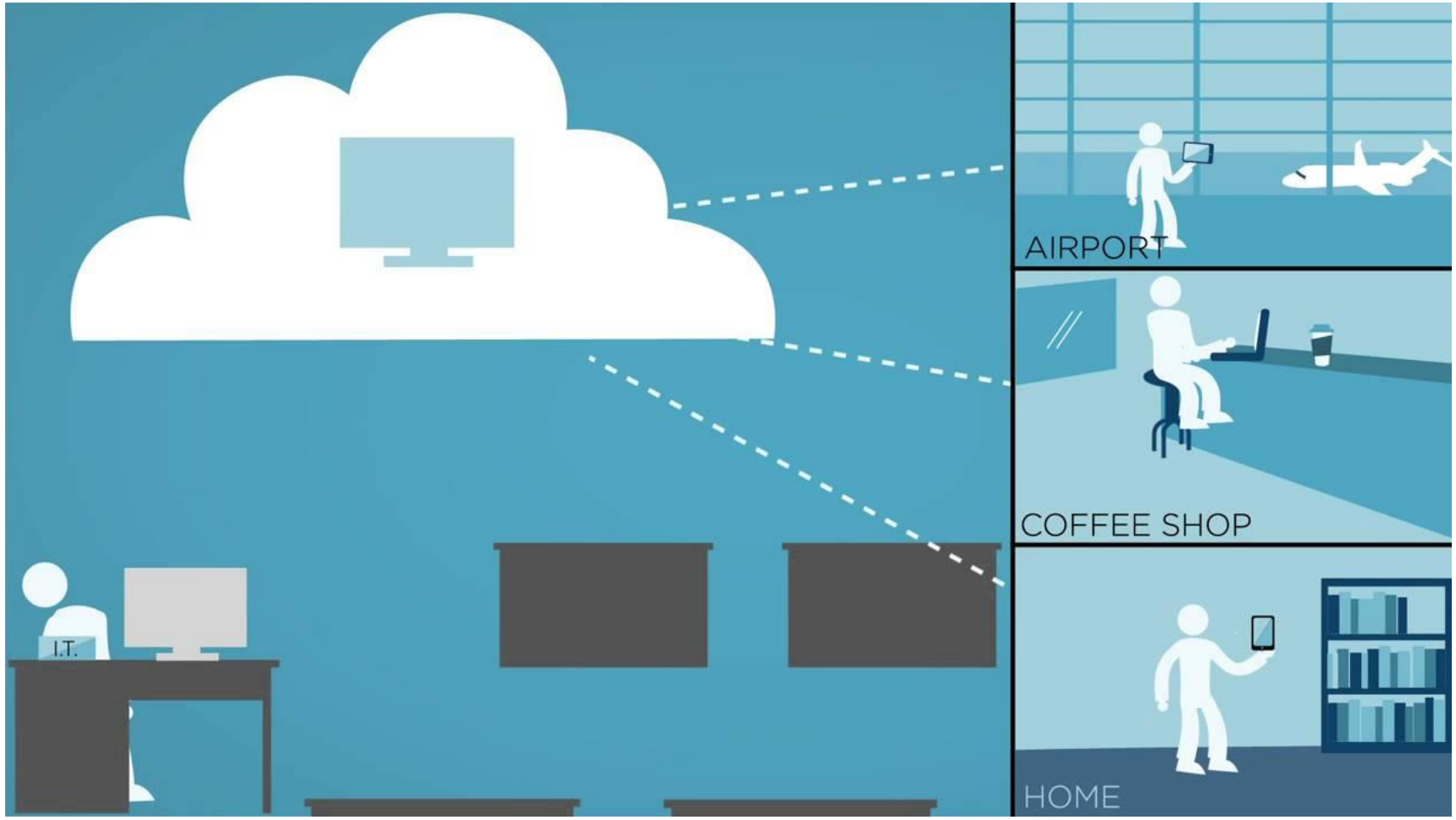
Scalability testing and VMware View Planner

Test results and Important takeaways

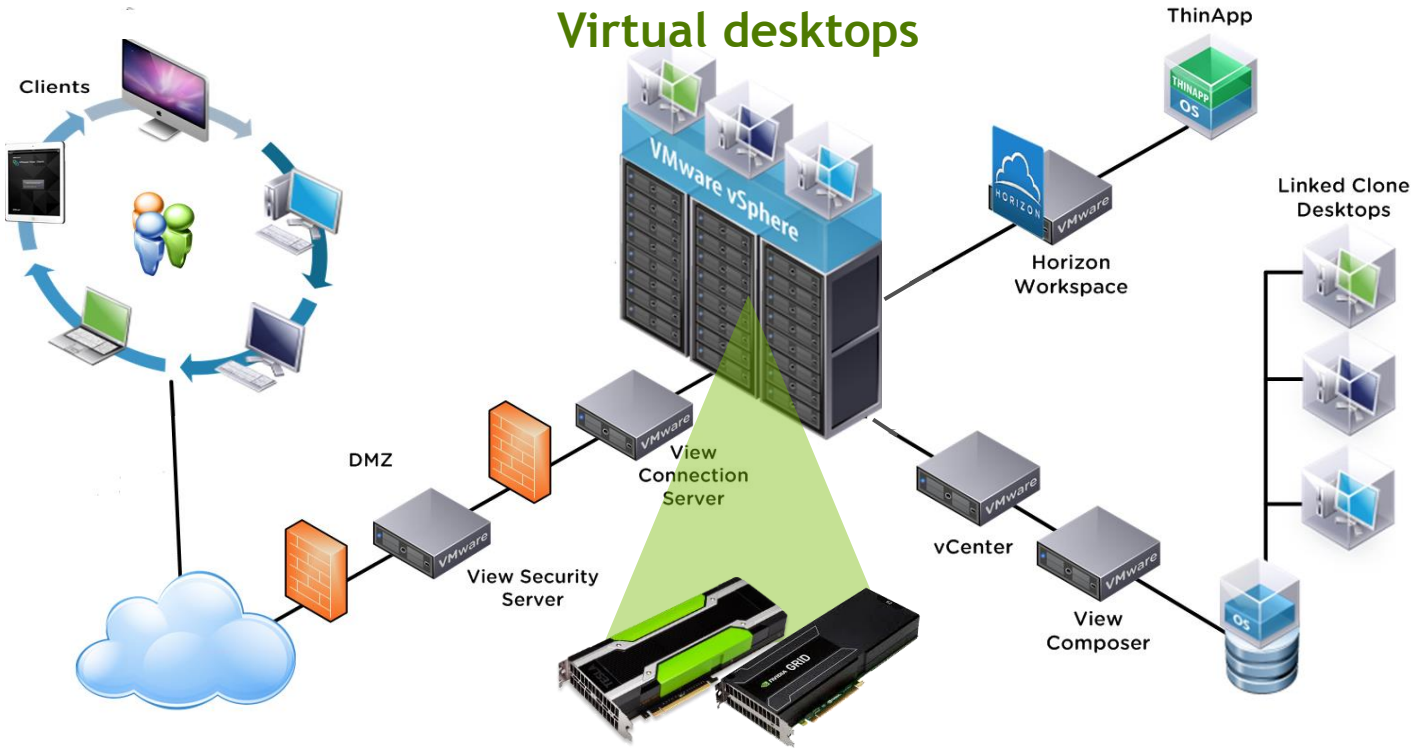
Best Practices

INTRODUCTION

VMWARE HORIZON VIEW OVERVIEW

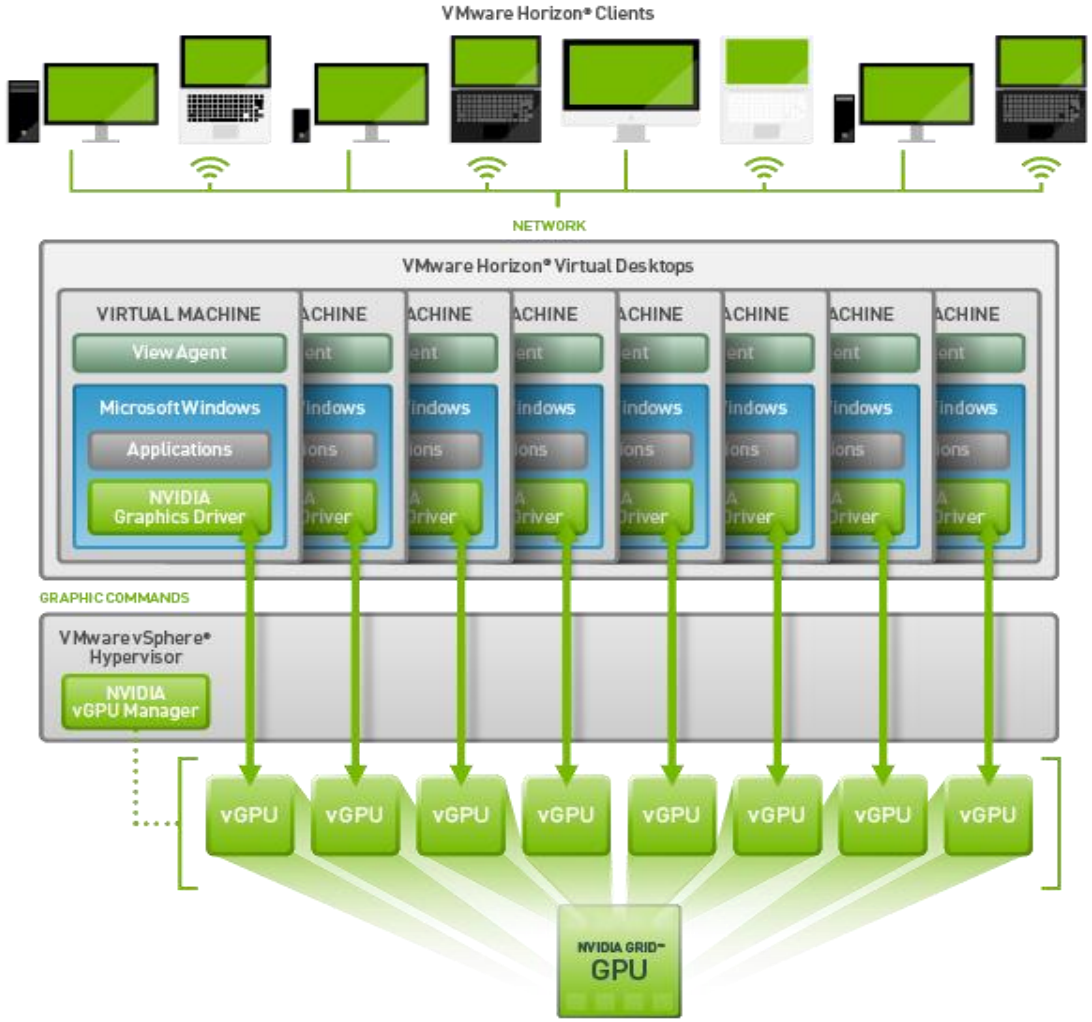


VMWARE HORIZON VIEW OVERVIEW

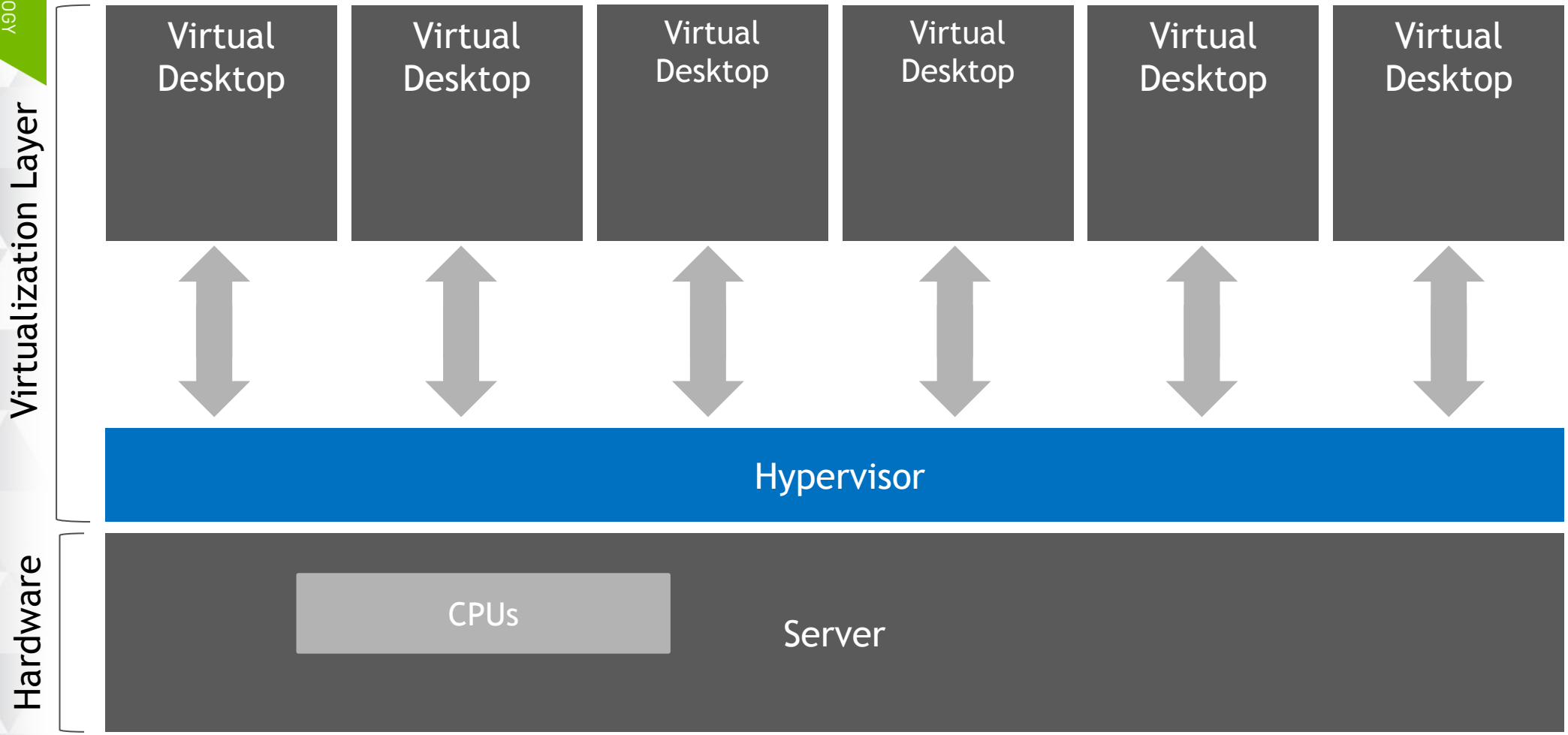


Enhancing performance and user experience with GPUs

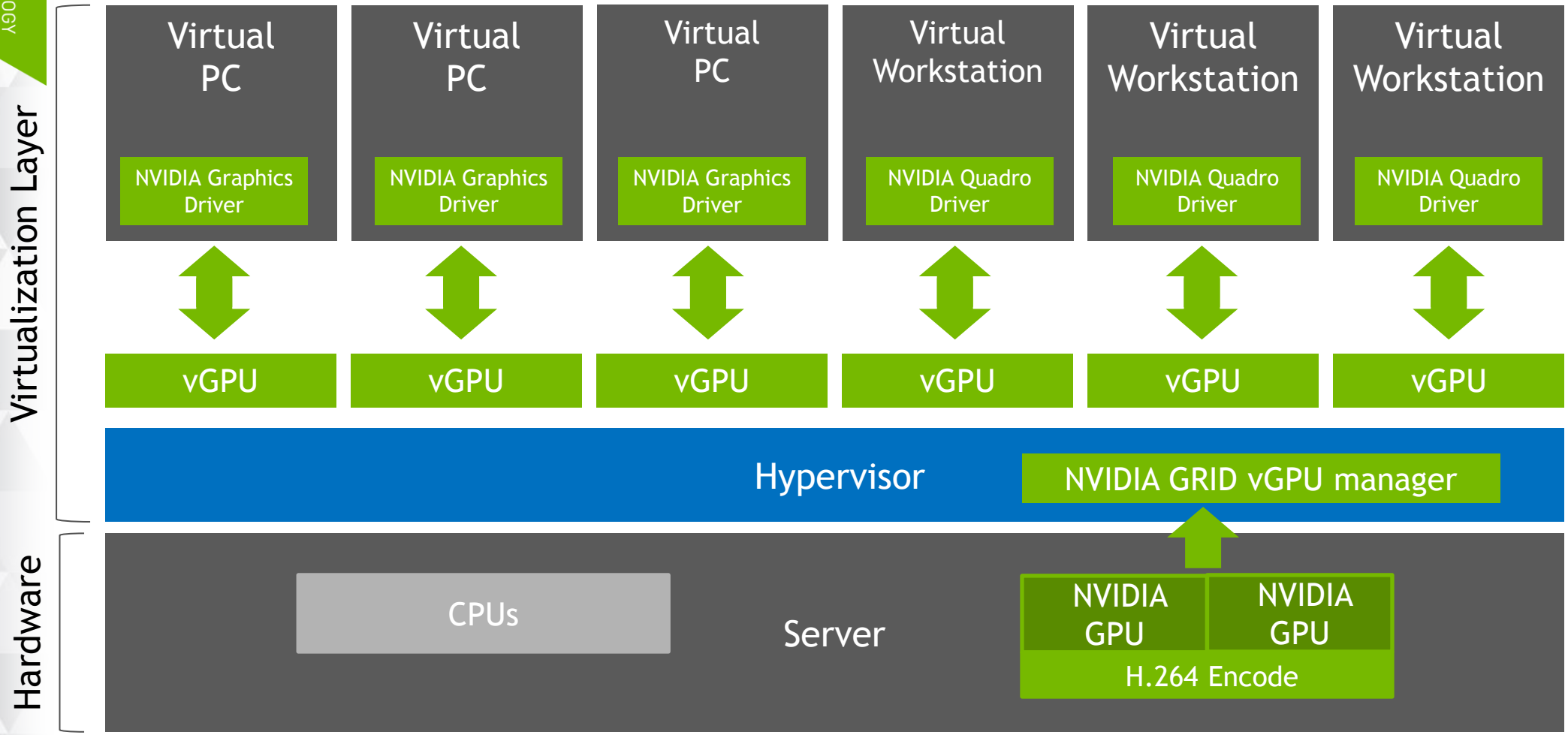
VMWARE HORIZON WITH NVIDIA GRID GPU



HOW DOES NVIDIA GRID WORK?



HOW DOES NVIDIA GRID WORK?



TIME SLICING

A Time Slice is the period of time for which a process is allowed to run in a preemptive multitasking system

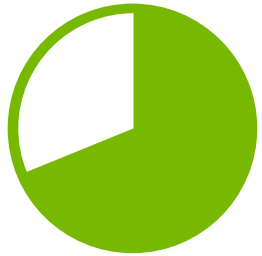
Time slicing is a leveraged by hypervisors (vSphere, XenServer, KVM, Hyper-V) to share physical resources (CPU, Network, I/O etc.) between multiple virtual machines

Time slicing allows the distribution of pooled resources based on actual need.

NVIDIA GRID uses time slicing to share the 3D engine between virtual machines


Getting lunch


In a meeting


Not in office


Thinking


Viewing information

Knowledge workers or engineers may be connected to virtual machines that share a physical GPU at the same time but typically don't utilize the physical GPU the entire time because human workflows include

During these times, the GPU isn't under load and can be shared with other virtual machines/users

NVIDIA VIRTUAL GPU TYPES

Card	Physical GPUs	GRID Virtual GPU	Intended Use Case	Frame Buffer (Mbytes)	Virtual Display Heads	Max Resolution per Display Head	Maximum vGPUs	
							Per GPU	Per Board
GRID K1	4	K180Q	Power User	4096	4	2560x1600	1	4
		K160Q	Power User	2048	4	2560x1600	2	8
		K140Q	Power User	1024	2	2560x1600	4	16
		K120Q	Power User	512	2	2560x1600	8	32
		K100	Knowledge Worker	256	2	1920x1200	8	32
GRID K2	2	K280Q	Designer	4096	4	2560x1600	1	2
		K260Q	Power User, Designer	2048	4	2560x1600	2	4
		K240Q	Power User, Designer	1024	2	2560x1600	4	8
		K220Q	Power User, Designer	512	2	2560x1600	8	16
		K200	Knowledge Worker	256	2	1920x1200	8	16

Table 1 GRID K1 and K2 Virtual GPU types

NVIDIA VIRTUAL GPU TYPES

Card	Physical GPUs	GRID Virtual GPU	Intended Use Case	Frame Buffer (Mbytes)	Virtual Display Heads	Max Resolution per Display Head	Maximum vGPUs	
							Per GPU	Per Board
Tesla M60	2	M60-8Q	Designer	8192	4	3840x2160	1	2
		M60-4Q	Designer	4096	4	3840x2160	2	4
		M60-2Q	Designer	2048	4	2560x1600	4	8
		M60-1Q	Power User, Designer	1024	2	2560x1600	8	16
		M60-0Q	Power User, Designer	512	2	2560x1600	16	32
		M60-2B	Power User	2048	2	2560x1600	4	8
		M60-1B	Power User	1024	2	2560x1600	8	16
		M60-0B	Power User	512	2	2560x1600	16	32
Tesla M6	1	M6-8Q	Designer	8192	4	3840x2160	1	1
		M6-4Q	Designer	4096	4	3840x2160	2	2
		M6-2Q	Designer	2048	4	2560x1600	4	4
		M6-1Q	Power User, Designer	1024	2	2560x1600	8	8
		M6-0Q	Power User, Designer	512	2	2560x1600	16	16
		M6-2B	Power User	2048	2	2560x1600	4	4
		M6-1B	Power User	1024	2	2560x1600	8	8
		M6-0B	Power User	512	2	2560x1600	16	16

Table 2 Tesla M60 and M6 Virtual GPU types

WHY BENCHMARK?

BENCHMARKING VIRTUALIZED ENVIRONMENTS

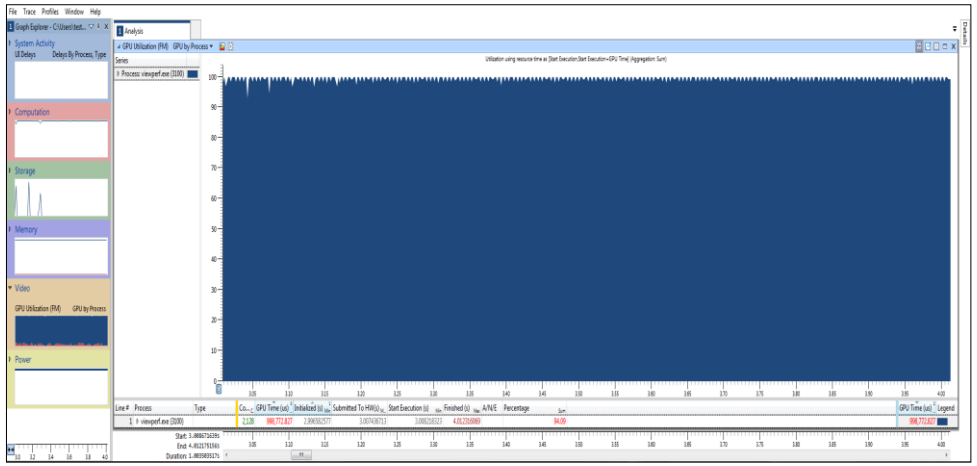
Typical Workstation benchmarks designed to stress all the available system resources.

Multiple VMs running the same task at the same time is not realistic test scenario

Most scalability tests can only simulate worst case real-user scenario

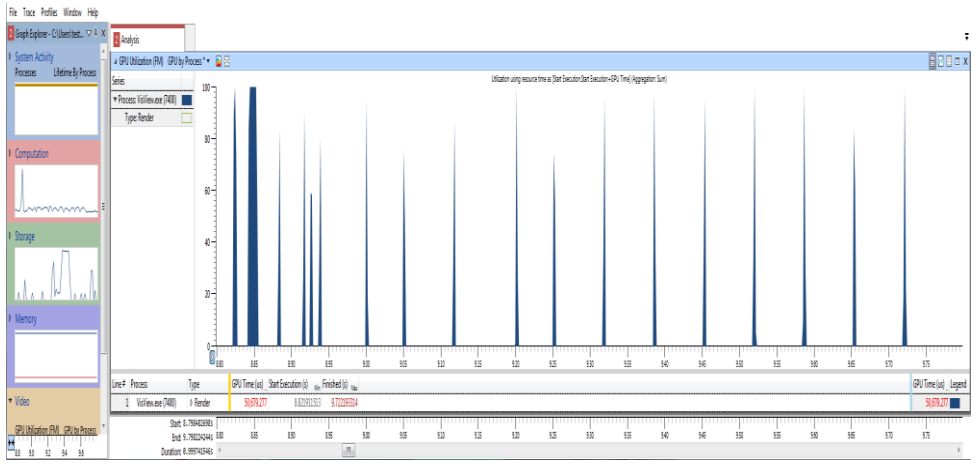
Benchmark

ViewPerf12 Catia viewset



Human workflow

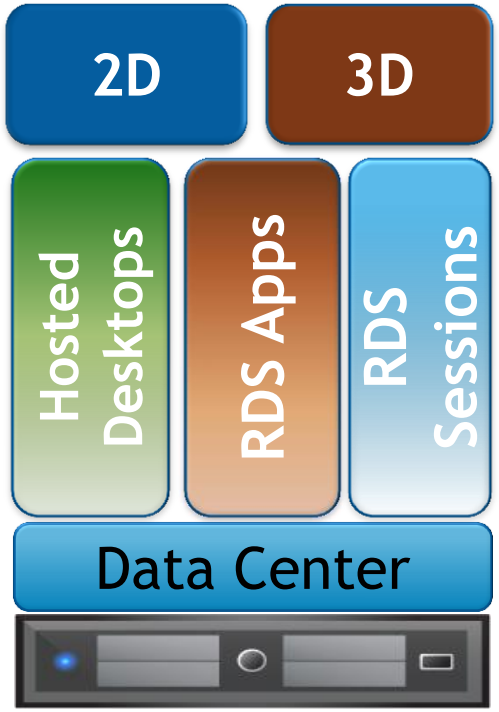
GPU ”heavy” process (zooming)



NEED

There is a need for

- End to end hardware/architecture comparison over generations
- Platform optimization and fine tuning
- ISV Certification process
- Sizing the VMs for best performance.
- Finding the right number of VMs that can run on the Host with acceptable performance
- Defining a workflow to automate the test process as the consolidation numbers and VM sizing will be different for different applications and physical hardware.



METHODOLOGY & TOOLS

HOW TO RUN SCALABILITY TESTS ?

- Ideal scenario would be testing with actual application users and monitoring the resource utilization over a extended period of time (days/weeks)
- LoginVSI Graphic workload
  - Multimedia Workload
  - Custom workload integration with LoginVSI
- VMware View Planner
  - Solidworks
  - 3DMark
  - Custom workload integration with View Planner
- In-house scripts for scalability test execution and log collection
  - AutoIT, Python, Powershell, psexec

PERFORMANCE METRICS AND USER EXPERIENCE

How to define a great User Experience ?

- Application FPS
- Application Response Time
- GPU statistics (nvidia-smi)
- Resource Utilization

- And more that needs to be defined

UX METRICS EXAMPLE

ESRI defined ArcGIS Pro UX based on following Performance Metrics:

Draw Time Sum - :80:90 seconds for basic tests to complete

Frame Per Second – 30-60 w/ 60 being optimal but ESRI admits 30 is ok, say users can't tell the difference

FPS Minimum – a big dip would mean the user saw a freeze, etc., below 5-10 FPS is an issue.

Standard Deviation – shows tests were uniform, quantity of tests:

<2 for 2D

<4 for 3D

WORKLOADS

Created and provided by ISV

- ESRI provided us with a 3D “heavy” workload: Philly 3D
- Adobe Photoshop graphics workload

Generic workloads for the ISV apps

- Revit RFO Benchmark
- Catalyst for AutoCAD

NVIDIA created workload

- AutoCAD workload was created by NVIDIA with help from AutoCAD

VMware View Planner workloads

SIZING

SIZING

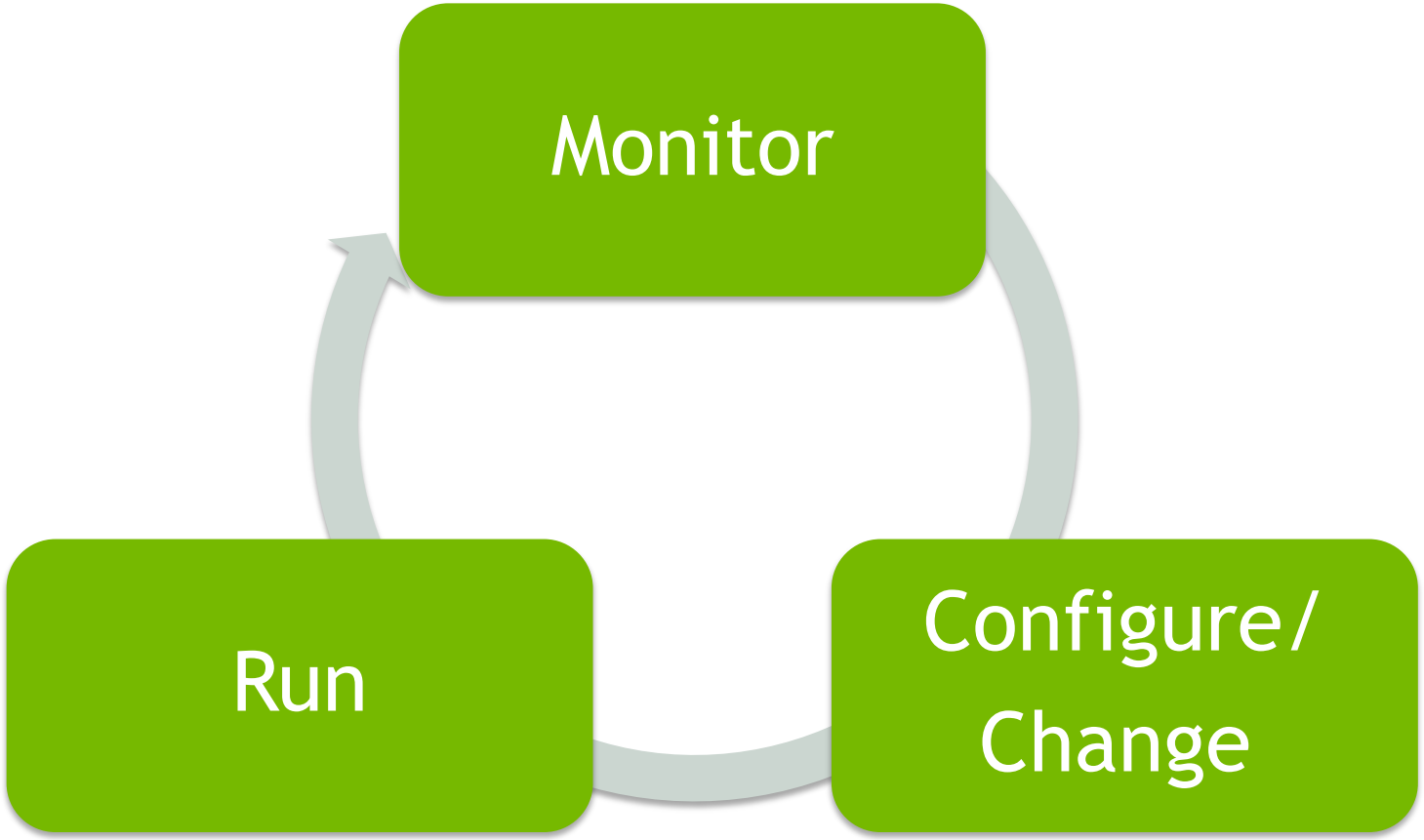
Use what you already know

- Size VM based on optimal physical workstation configurations
- Select vGPU profile based on Frame buffer requirements
- Apply all hypervisor recommended best practices
- Monitor VM resource utilization for a single VM test
- change VM resources based on the Max resource utilization

Important not to over-allocate VM resources for virtualized environment

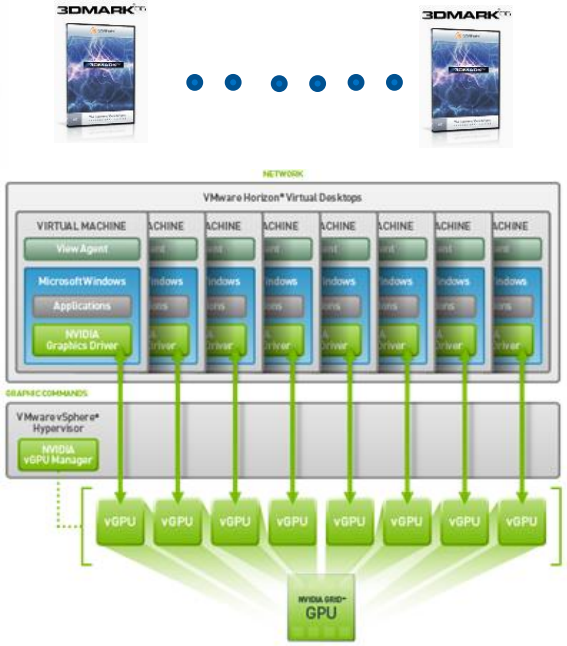
- Resource over allocation can reduce the performance of a VM as well as other VMs sharing the same host.
- Disabling hardware devices (typically done in BIOS) can free interrupt resources

SIZING METHODOLOGY

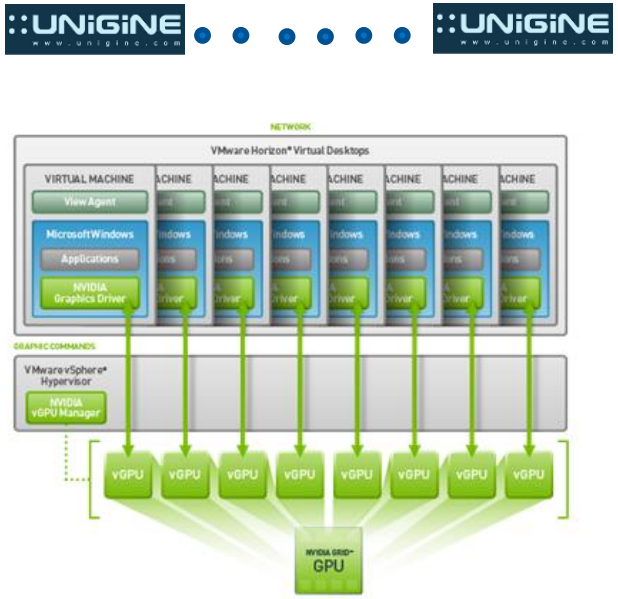


SIZING METHODOLOGY

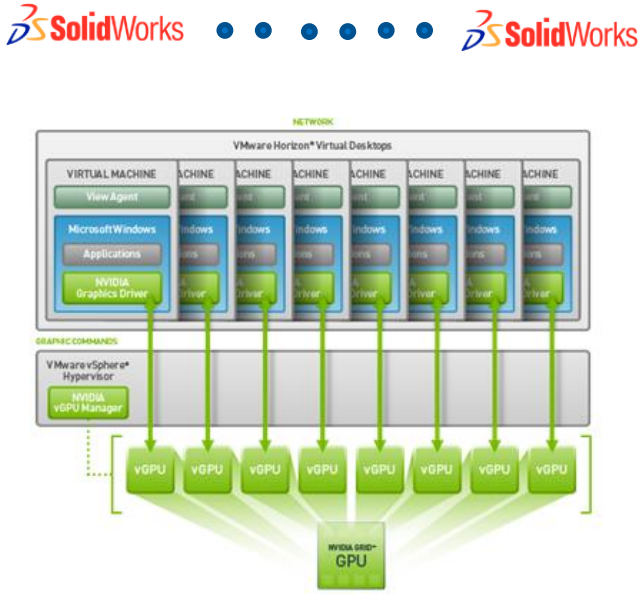
"x" users



"y" users



"z" users



VIEW PLANNER WORKLOAD

- View Planer allows benchmarking:
  - A mix of regular workload and any applications you bring
    - Health care, education, 3D graphics, Office 365
  - Up to thousands of VDI desktops or more

For Windows

Regular Workload

MS Office

Other Apps

3D Workload

SolidWorks

AUTODESK.

SIEMENS

3DMARK06

CAD Viewer 9.0

UNIGINE

VIEW PLANNER WORKLOAD

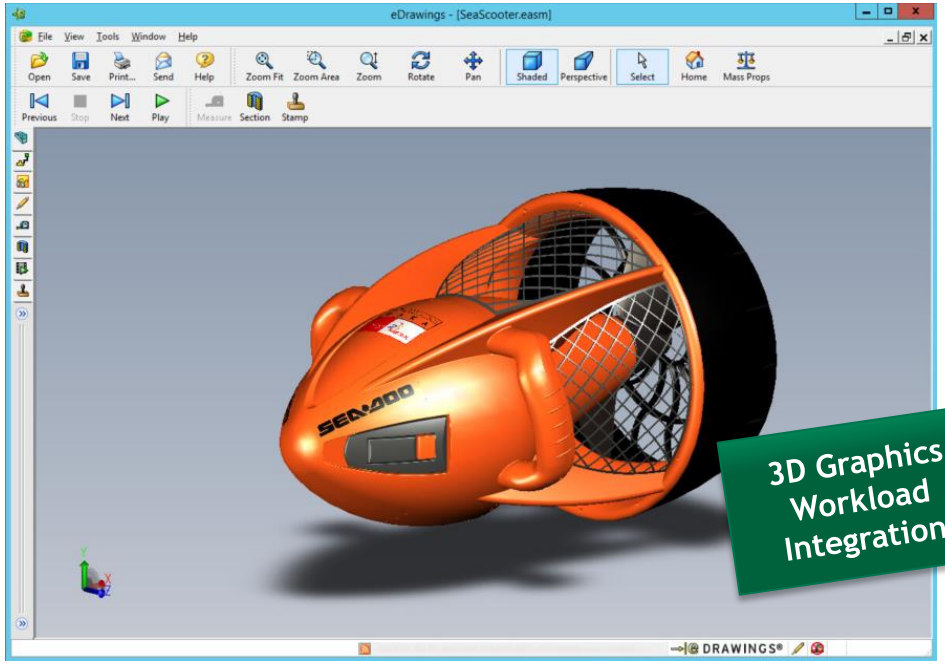
- View Planer allows benchmarking:
  - A mix of regular workload and any applications you bring
    - Health care, education, 3D graphics, Office 365
  - Up to thousands of VDI desktops or more

For Linux

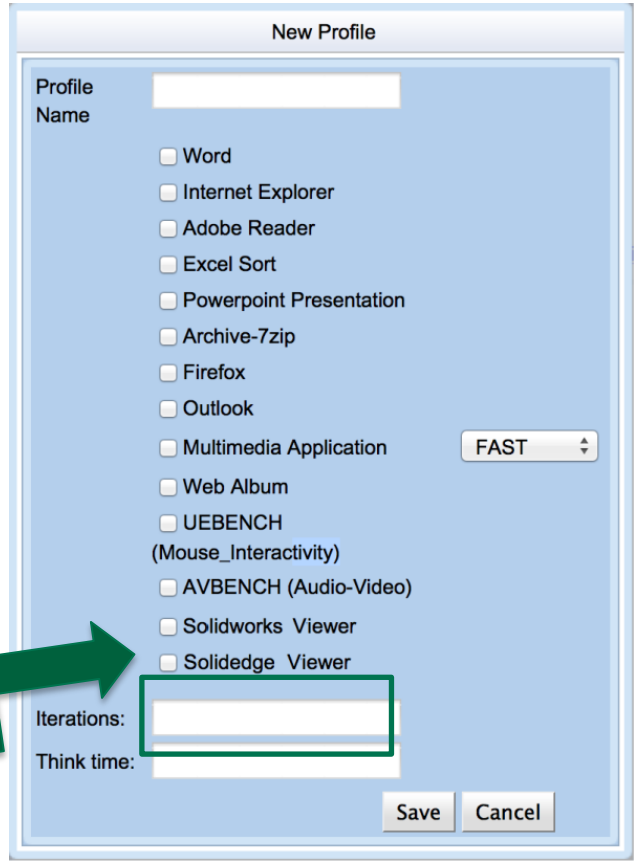
Regular Workload

BRING YOUR OWN APPLICATIONS (BYOA)

- Add your own customized applications
  - Including 3D graphics ones



3D Graphics Workload Integration



BENCHMARKING 3D GRAPHICS WORKLOAD WITH GRID VGPU

- Set up **Horizon View** and **View Planner**
- Set up **GRID vGPU** for **VDI Desktop / RDSH Server** Virtual Machine
- Register **3D Applications** with **View Planner**
- Create **Workload Run Profile** & start **Benchmarking Run**
- Collect **Benchmarking Results**

Run at Scale

SCALABILITY TESTS

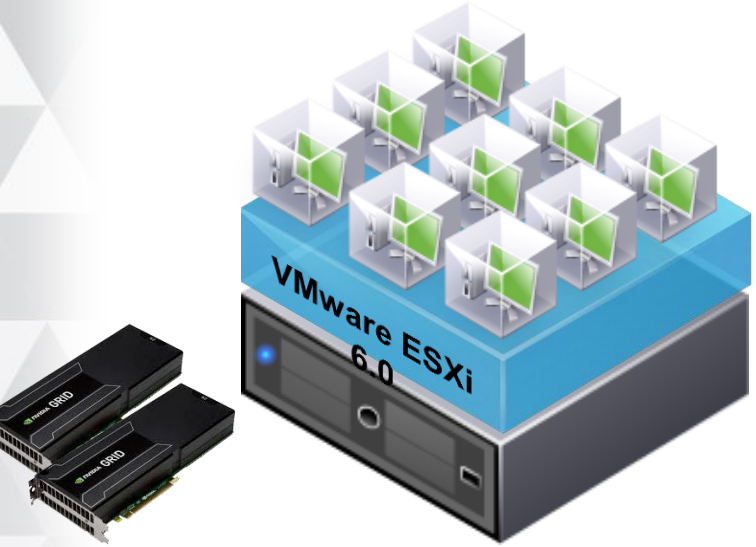
NVIDIA TEST SETUP

Virtual Client VMs

- 64-bit Win7 (SP1)
- 4vCPU, 4 GB RAM
- View Client 4.0

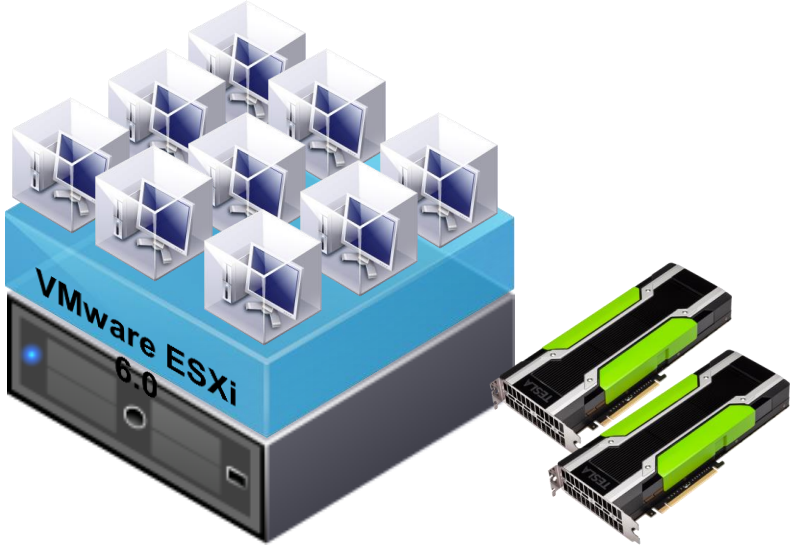
Virtual VDI desktop VMs

- 64-bit Win7 (SP1)
- 6vCPU, 14 GB RAM, 50GB HD
- Horizon View 7.0 agent



Remote Display Protocol

Blast Extreme / PCoIP



SuperMicro SYS-2027GR-TRFH
Intel Xeon E5- 2690 v2 @ 3.00GHz + 2 x Nvidia GRID K1
20 cores (2 x 10-core socket) Intel IvyBridge
256 GB RAM

SuperMicro SYS-2028GR-TRT
Intel Xeon E5-2698 v3 @ 2.30GHz + 2 x Nvidia GRID M60
32 cores (2 x 16-core socket) Intel Haswell
256 GB RAM

SPEC VIEWPERF 12 OVERVIEW

Workstation benchmark that can be used to simulate different workstation application workloads.

Tests the worst case scenario for virtualized and shared resource configuration.

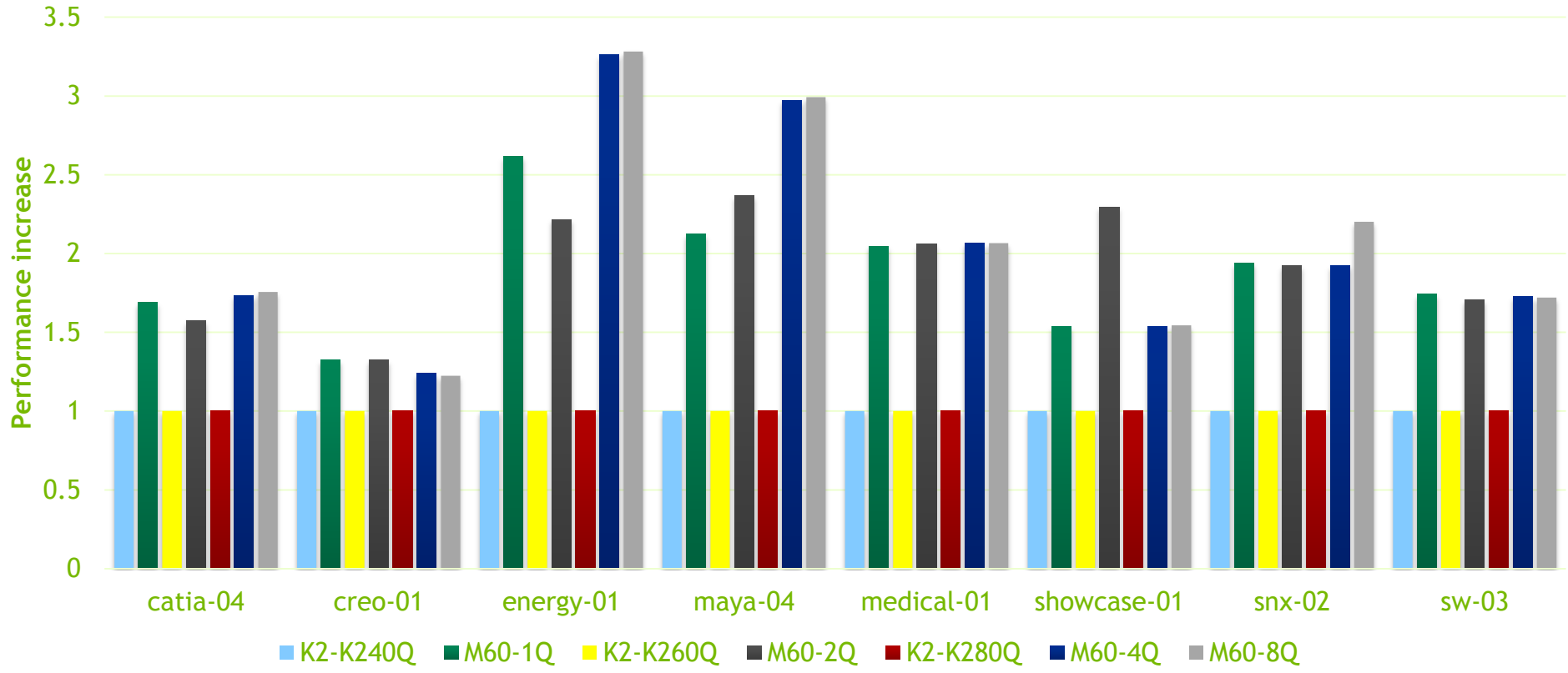
Can be used to Compare relative performance difference between:

- vGPU boards (K2 vs M60).
- Various remoting Protocols (PCoIP vs Blast) and the affect on the application performance.
- Can still be used for ISV certification process

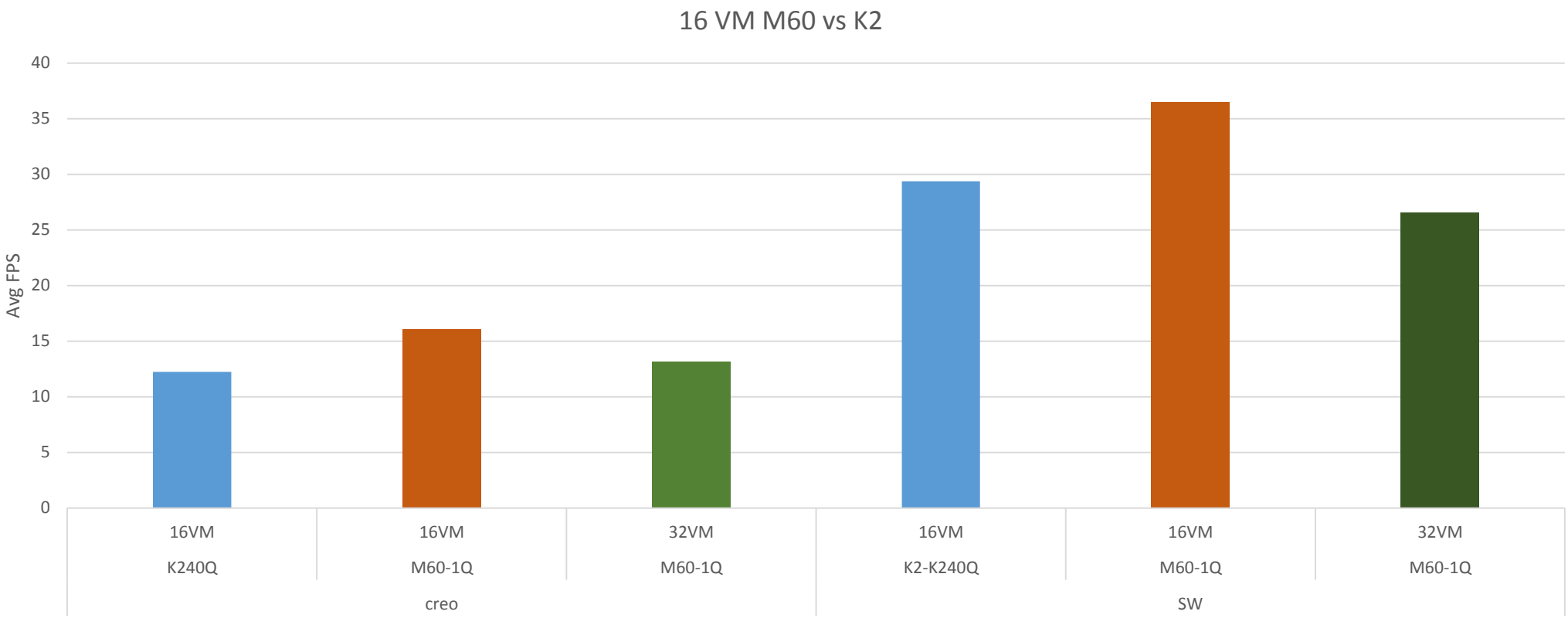
Cannot be used for sizing the VMs/Hosts.

SINGLE VM K2 VS M60 VIEWPERF 12

Increase in Performance
4vcpu / 14GB RAM

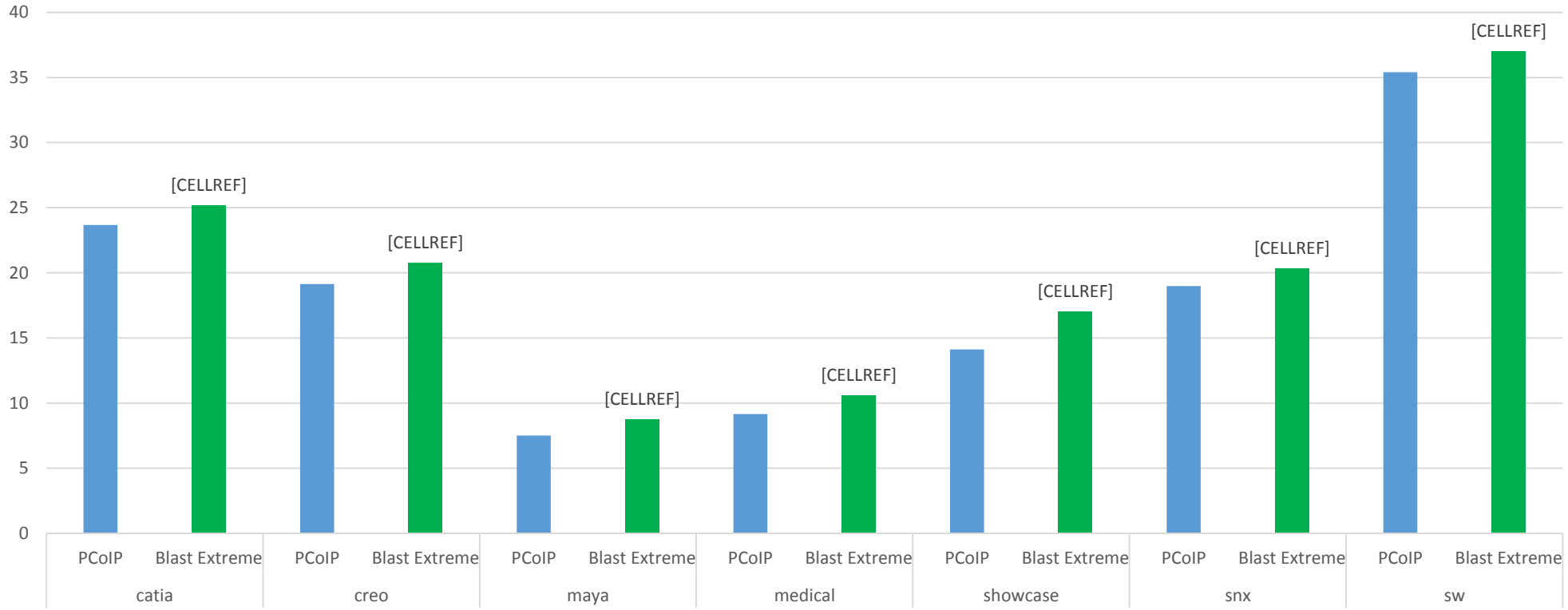


VIEWPERF12 SCALABILITY TESTS



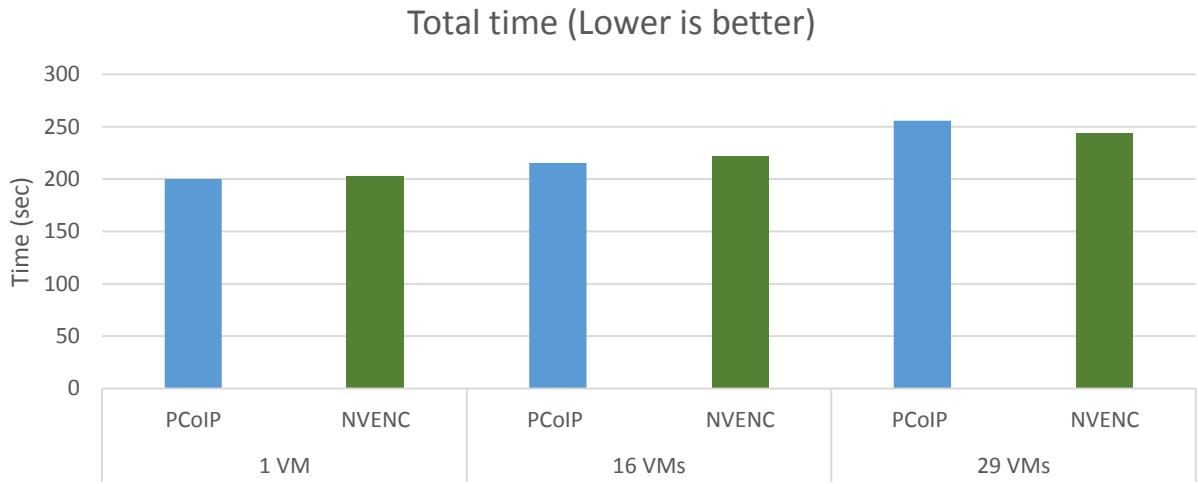
VIEWPERF12 SCALABILITY TEST

Comparing Remoting Protocols
16 VMs test (4vCPU/14GB RAM/M60_2Q)



REVIT RFOBENCHMARK

6vCPU 6GB RAM M60_1Q	Model creation and view export benchmark										CPU Render benchmark	Graphics benchmark w/ hardware acceleration				Graphics benchmark w/o hardware acceleration			
	opening and loading the custom template	creating the floors levels and grids	creating a group of walls and doors	modifying the group by adding a curtain wall	creating the exterior curtain wall	creating the sections	changing the curtain wall panel type	export all views as PNGs	export some views as DWGs	Total	render w/ nvidia mental ray	refresh Hidden Line view x12	refresh Consistent Colors view x12	refresh Realistic view x12	rotate view x1	refresh Hidden Line view x12	refresh Consistent Colors view x12	refresh Realistic view x12	rotate view x1
PCoIP 1 VM	3.73	12.83	27.08	54.07	15.10	9.38	5.07	33.07	39.71	200.04	117.68	6.93	6.65	8.75	2.25	27.48	28.29	31.84	11.04
Blast 1 VM	5.18	12.84	26.98	55.42	14.91	9.56	5.09	32.70	40.01	202.69	119.65	7.12	6.71	8.92	2.27	27.93	28.23	31.30	11.07
PCoIP 16 VMs	4.24	13.06	28.55	57.90	15.68	9.70	5.27	36.51	44.55	215.45	296.10	7.65	7.40	9.87	2.45	38.59	41.47	45.32	15.48
Blast 16 VMs	4.02	13.13	29.08	58.61	15.63	9.82	5.32	37.26	45.23	222.42	303.50	7.57	7.42	9.99	2.52	40.11	42.44	43.90	14.38
PCoIP 29 VMs	4.42	14.72	32.92	67.19	17.41	10.70	5.80	45.88	56.65	255.70	554.45	8.66	8.37	11.10	2.69	71.55	82.02	76.86	23.40
Blast 29 VMs	4.33	14.31	31.35	64.56	17.39	10.41	5.74	41.47	54.06	243.62	536.20	9.21	8.67	11.62	2.80	68.31	78.72	77.53	27.61

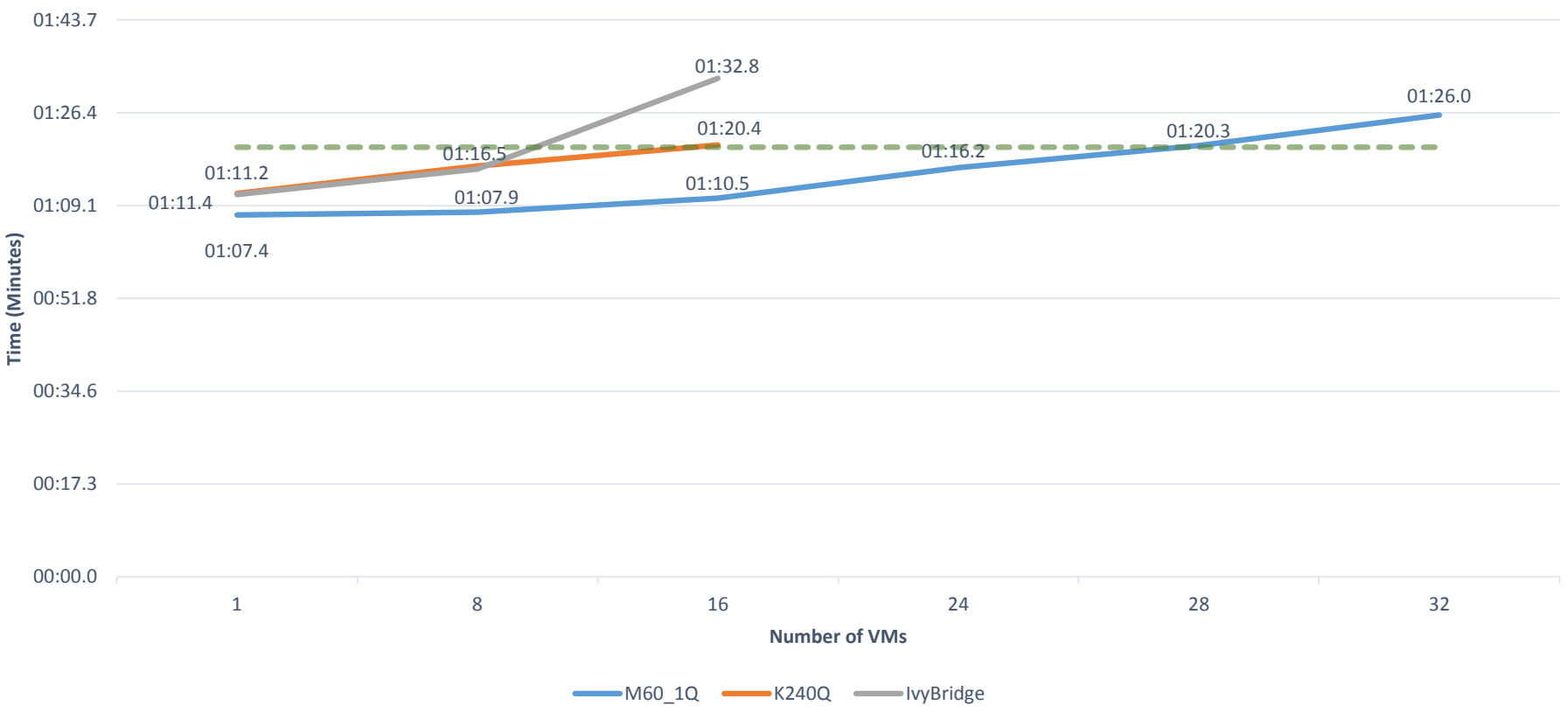


ESRI ARCGIS PRO 10.0 TEST RESULTS

	Application Metrics						GPU		CPU %Core Util	
	VM Config	VM count	DrawTime (min:sec)	FPS	Min FPS	Std deviation	%Util	%Mem	Avg	Max
Philly 3D Map Think Time 5 Seconds Navigation Time 5 Second	Intel Ivy Bridge K240q 6vcpu 6GB RAM	1	01:11.2	62.34	21.96		12	2	13	22
		8	01:15.9	53.48	14.14	1.9	43	9.5	62.885	95.55
		12	01:22.6	45.32	8.65	2.9	66	8.3	74	98.43
		16	01:32.8	40.7	6	3.7	57	12	95.786	99.99
	Haswell K240q 6vcpu 6GB RAM	1	01:11.4	65.85	35.41		7	2	9.5	19.62
		8	01:16.5	60.92	27.61	1.03	52	10.3	50.17	67.3
		12	01:17.3	55.25	19.05	3.8	54	10.7	57.53	84.15
		16	01:20.4	47.28	13.4	2.25	63	12	67.27	94.07
	Haswell **M60-1Q** 6vcpu 6GB RAM	1	01:07.4	66.52	42.74		8	2	7.9	12
		8	01:07.9	63.43	34.91	0.34	44	7	27.557	38.37
		16	01:10.5	57.74	24.96	0.82	71	12	50.145	65.34
		24	01:16.2	50.85	16.99	3.03	92	28	69.316	81.24
		28	01:20.3	47.54	13.81	3.90	96	28	75.41	84.64
		32	01:26.0	43.42	11.37	5.7	94	20	78.52	88.59

ESRI ARCGIS PRO 10.0 DRAW-TIME SUM

As



ESRI ArcGIS Pro 3D

UPH – Users per Host

ESRI Heavy 3D Workload

12 UPH

K240Q Users
6vCPU – 6GB RAM

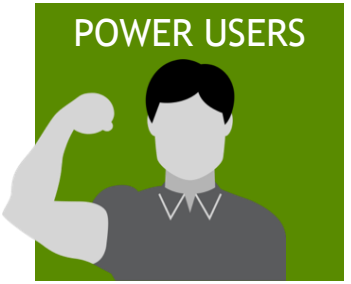
2x NVIDIA GRID K2

Medium 3D Workload

16 UPH

K240Q Users
6vCPU – 6GB RAM

Lab host:
CPU: Dual Socket 2.3Ghz / 16 core
RAM: 256GB RAM
GPU: 2 NVIDIA GRID K2 cards
10G Core network
iSCSI SAN: ~25K max IOPS
VMware vSphere 6
VMware Horizon 6.1 w/ vGPU
Tested 6/2015



ESRI ArcGIS Pro 3D

UPH – Users per Host

ESRI Heavy 3D Workload

24 UPH

M60_1Q Users
6vCPU – 6GB RAM

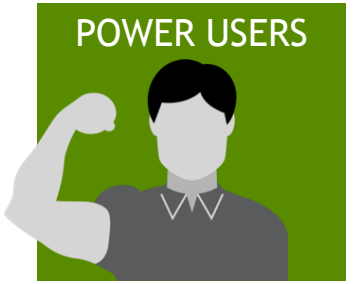
2x NVIDIA GRID M60

Medium 3D Workload

28 UPH

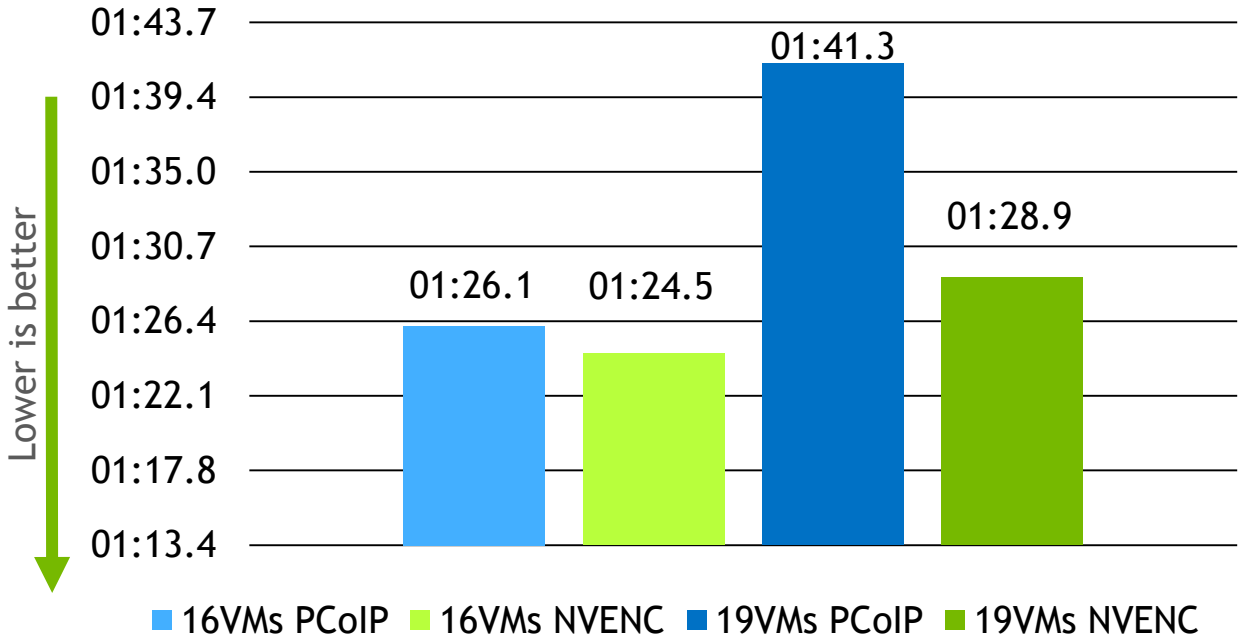
M60_1Q Users
6vCPU – 6GB RAM

Lab host:
CPU: Dual Socket 2.3Ghz / 16 core
RAM: 256GB RAM
GPU: 2 NVIDIA GRID M60 cards
10G Core network
iSCSI SAN: ~25K max IOPS
VMware vSphere 6
VMware Horizon 6.1 w/ vGPU
Tested 6/2015



ARCGIS PRO 10.2 SCALABILITY TEST

- 16 (M60_2Q) and 19 (1Q) VMs running ESRI ArcGIS Pro.
- Draw Time of around 01:20 minutest guarantees a great user experience.

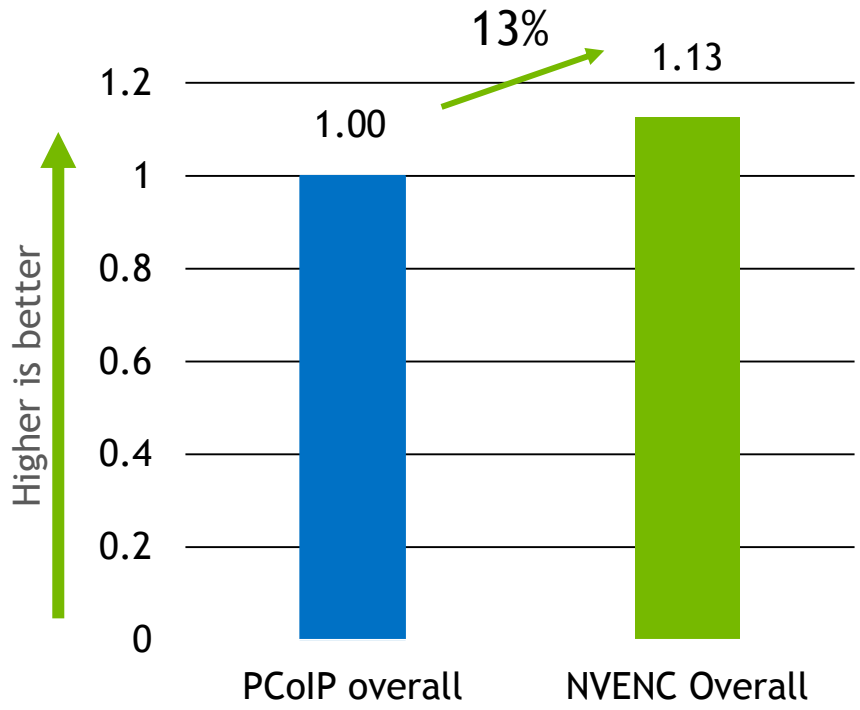


Protocol acceleration increases users per host by 18% (3VMs) for ESRI ArcGIS Pro 1.1 3D users

Source: NVIDIA GRID Performance Engineering Lab

INCREASES AVERAGE FPS

VP 12 PCoIP vs Blast Overall

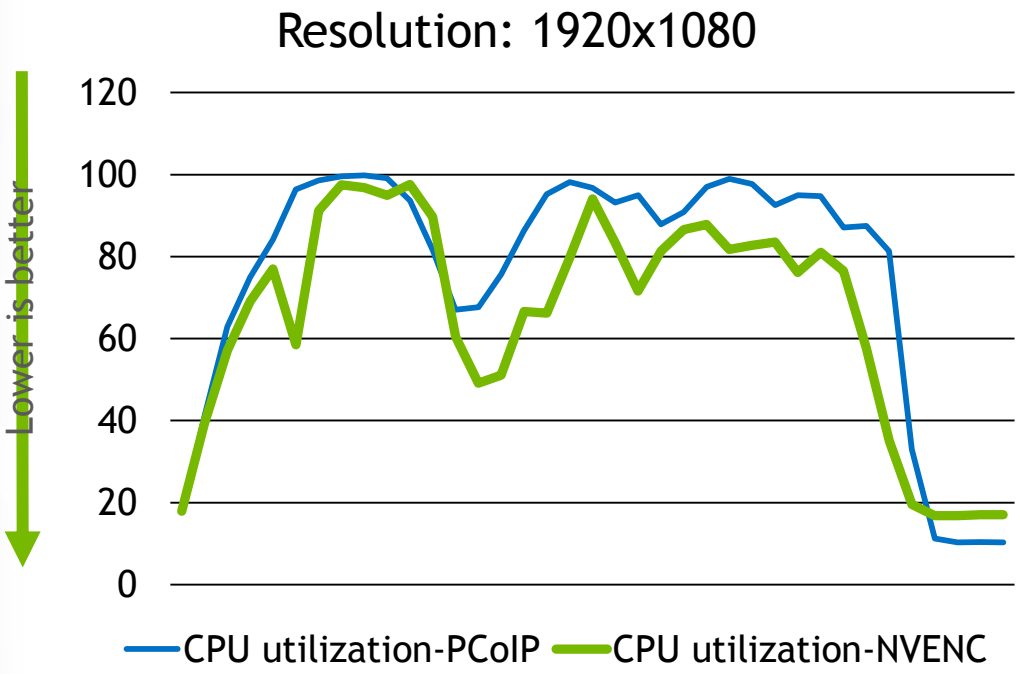


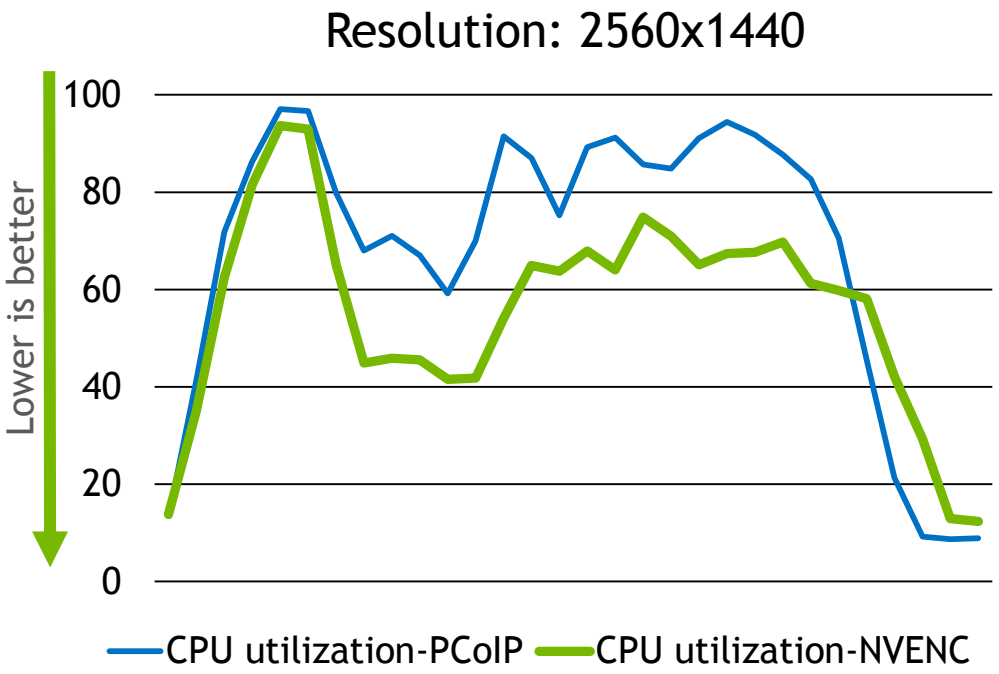
Protocol acceleration increases average FPS by 13% across VP12 subtests running 16 VMs (2Q). Dependent on subtests the performance difference varies between -2.02% and 25.27%

Source: NVIDIA GRID Performance Engineering Lab

DECREASES CPU LOAD

Host CPU utilization of 19 VMs (1Q) running ESRI ArcGIS Pro.





Impact of protocol acceleration increases with the amount of pixels. 16% (1920x1080) -> 22% (2560x1440).

Source: NVIDIA GRID Performance Engineering Lab

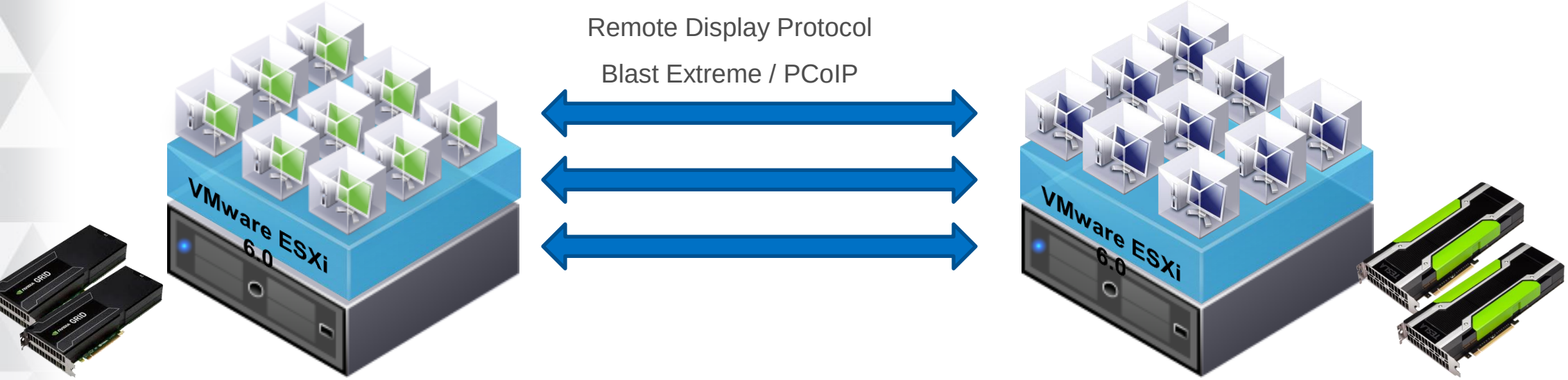
VIEW PLANNER TESTBED

Virtual Client VMs

- 64-bit Win7 (SP1)
- 4vCPU, 4 GB RAM
- View Client 4.0

Virtual VDI desktop VMs

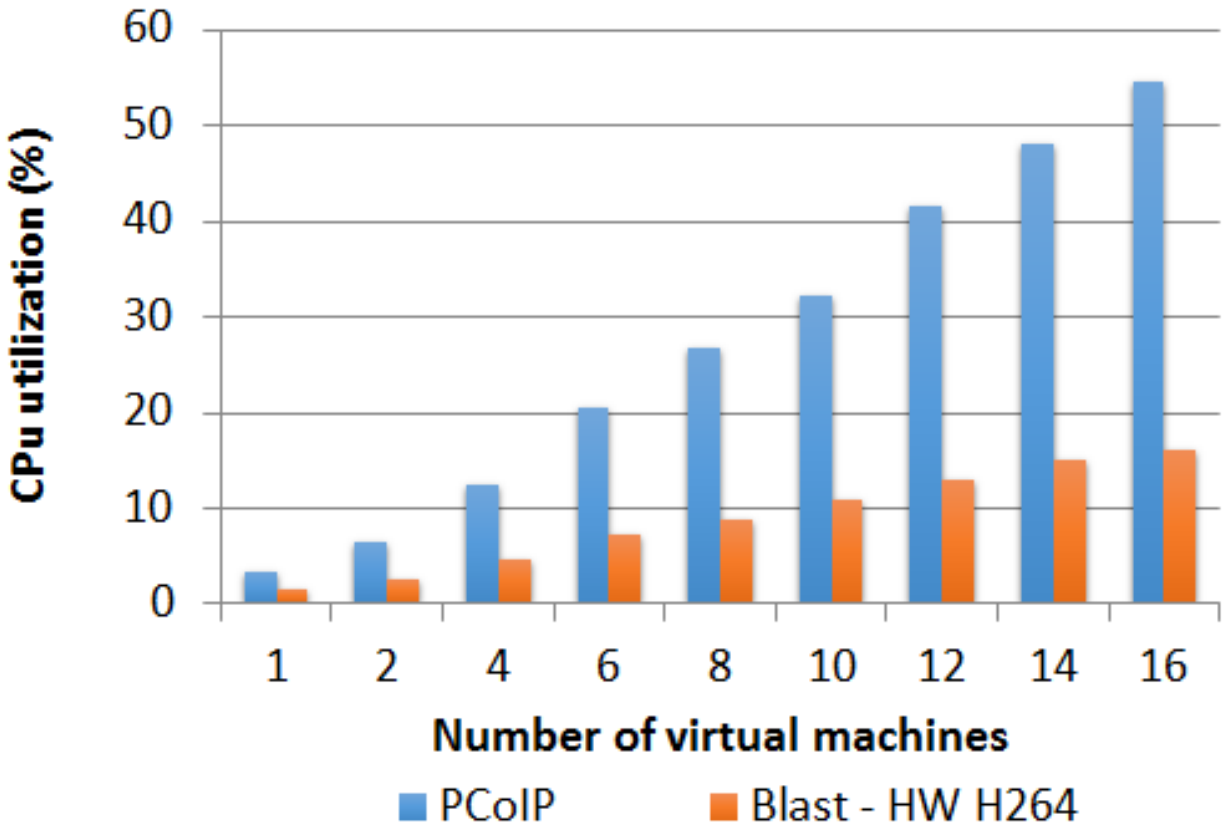
- 64-bit Win7 (SP1)
- 4vCPU, 20 GB RAM, 24GB HD
- Horizon View 7.0 agent



Dell R730 – Intel Haswell CPUs + 2 x Nvidia GRID K1
24 cores (2 x 12-core socket) E5-2680 V3
384 GB RAM

Dell R730 – Intel Haswell CPUs + 2 x Nvidia GRID M60
24 cores (2 x 12-core socket) E5-2680 V3
384 GB RAM

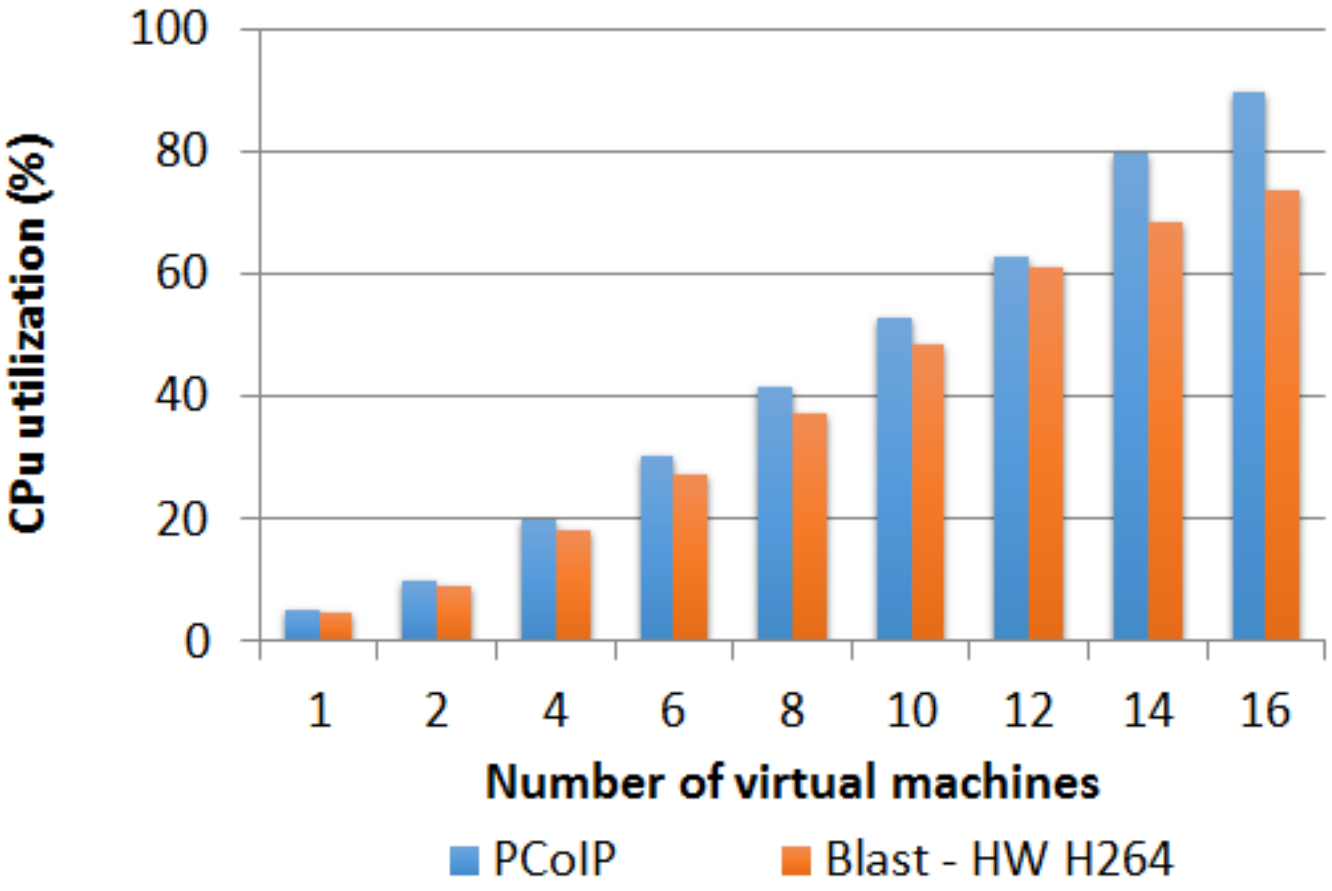
SPECapc for 3DSmax - CPU utilization at the client



SPECapc for 3DSmax - CPU utilization at the client

# of VM	PCoIP (A)	Blast - HW H264 (B)	A / B
1	3 %	2 %	2.2 x
2	6 %	2 %	2.6 x
4	12 %	5 %	2.7 x
6	20 %	7 %	2.9 x
8	27 %	9 %	3.1 x
10	32 %	11 %	3.0 x
12	41 %	13 %	3.2 x
14	48 %	15 %	3.2 x
16	54 %	16 %	3.4 x

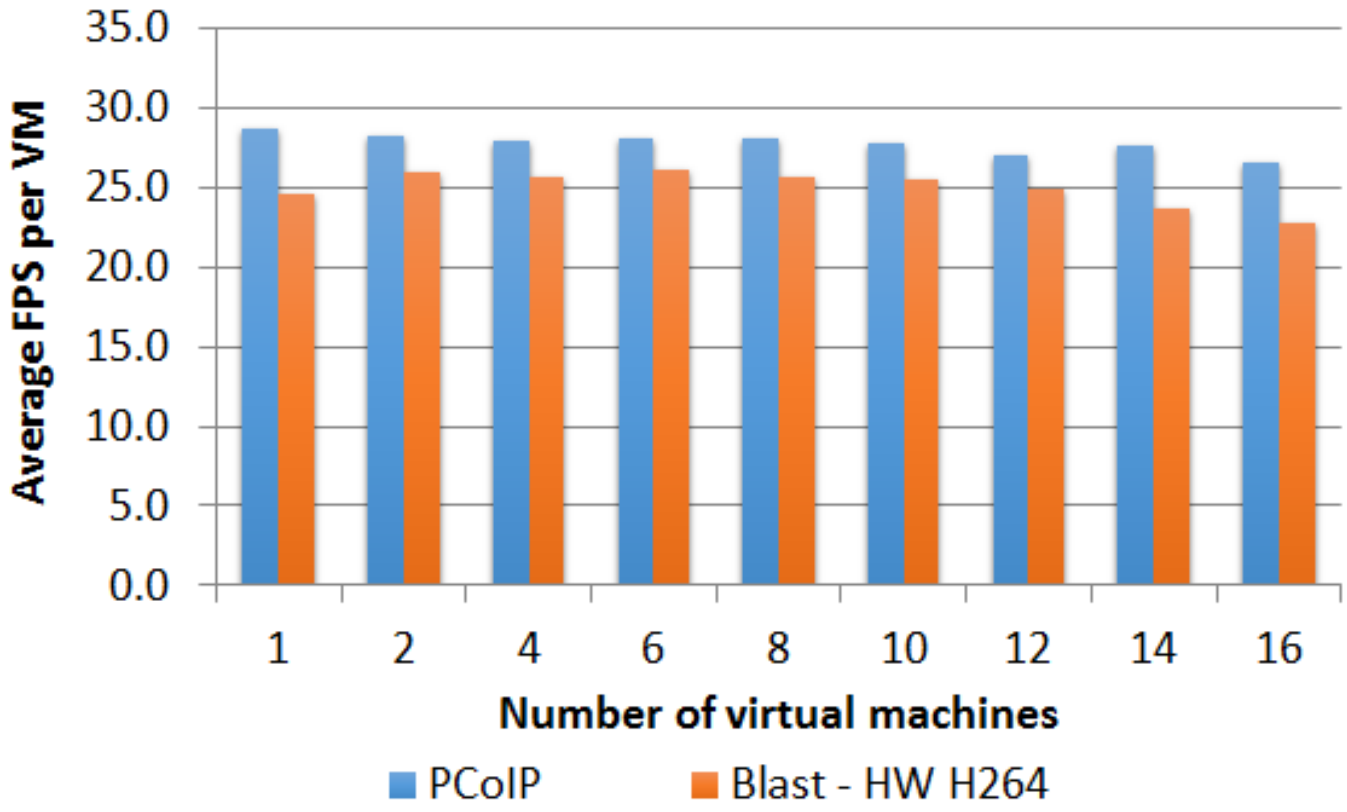
SPECapc for 3DSmax - CPU utilization at the server



SPECapc for 3DSmax - CPU utilization at the client

# of VM	PCoIP (A)	Blast - HW H264 (B)	A / B
1	5 %	5 %	1.1 x
2	10 %	9 %	1.1 x
4	20 %	18 %	1.1 x
6	30 %	27 %	1.1 x
8	42 %	37 %	1.1 x
10	53 %	48 %	1.1 x
12	63 %	61 %	1.03 x
14	80 %	68 %	1.2 x
16	90 %	74 %	1.2 x

SPECapc for 3DSMax 2015 – Average FPS per VM delivered at the client



BEST PRACTICES

BEST PRACTICES

We have seen Host Turboboost setting greatly impact performance. Evaluate with or without this Bios setting for your use case.

For the Host CPU, We have seen that the higher number of cores impact scalability more than higher clock speed.

Consider distributing the VMs evenly across all the GPUs

Try to size the VM within the NUMA node boundaries.

Proper single VM sizing very important for higher scalability.

GPU TECHNOLOGY CONFERENCE

April 4-7, 2016 | Silicon Valley

THANK YOU

JOIN THE CONVERSATION

#GTC16

JOIN THE NVIDIA DEVELOPER PROGRAM AT developer.nvidia.com/join

PRESENTED BY

NVIDIA.