



OPTIMIZING NVIDIA VIRTUAL GPU FOR THE BEST VDI USER EXPERIENCE

NVIDIA VIRTUAL GPU PRODUCT POSITIONING

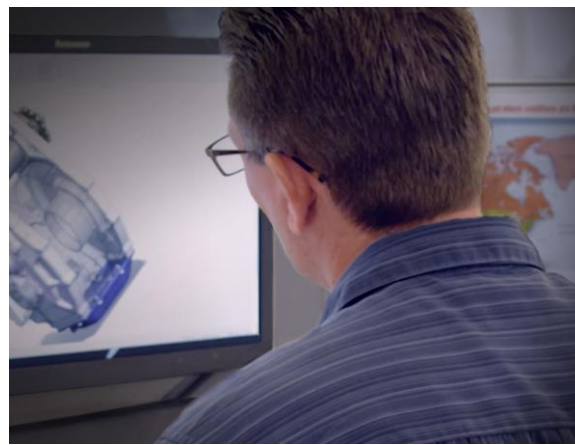
NVIDIA GRID vPC/vApps



Knowledge/Business
Worker

Tesla M10

NVIDIA QUADRO Virtual Data Center Workstation



Engineers/ Architects/
Designers

Tesla P4*

* Exception High End and Ultra High-End Use Cases

GRID vPC and Quadro vDWS

Understanding the workflow to define scale

NVIDIA GRID vPC/vApps

Scale determined by
Framebuffer Size*

All Maxwell and Pascal based
Tesla boards provide sufficient
3D Performance for typical
GRID vPC workloads

	Tesla M10 (8GB) 8 Users	Tesla P40 (24GB) 24 Users
End-User Latency	~200ms	~200ms
Frames/User	4000	4000

NVIDIA QUADRO Virtual Data Center Workstation

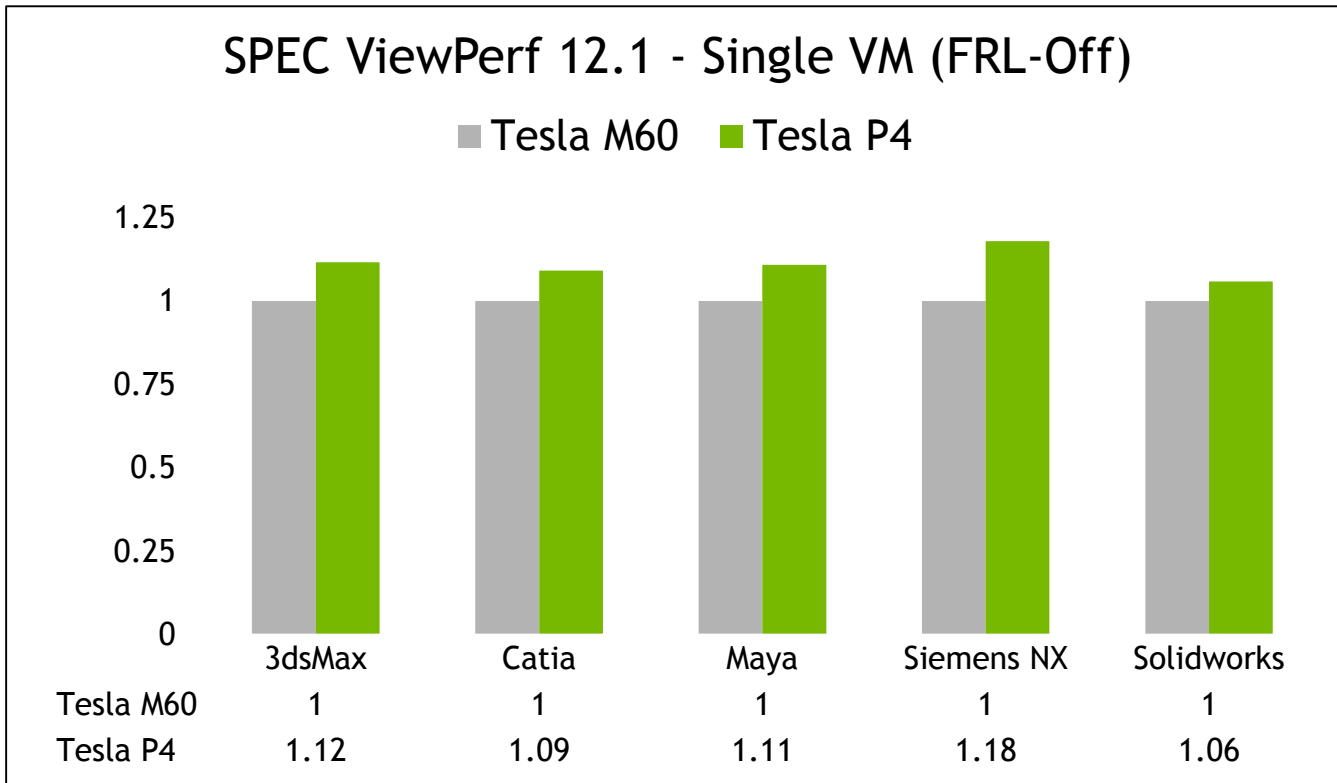
Scale determined by
3D Engine Performance and
Framebuffer Size*

	Tesla M10 (8GB) 1 User	Tesla P4 (8GB) 1 User
SPEC ViewPerf 12.1	~25	~80

* Tested with Single Full HD Screen. Subject to change with non Pascal and Volta based GPUs

QUADRO Virtual Data Center Workstation

P4 provides 11% more Perf than each M60 GPU

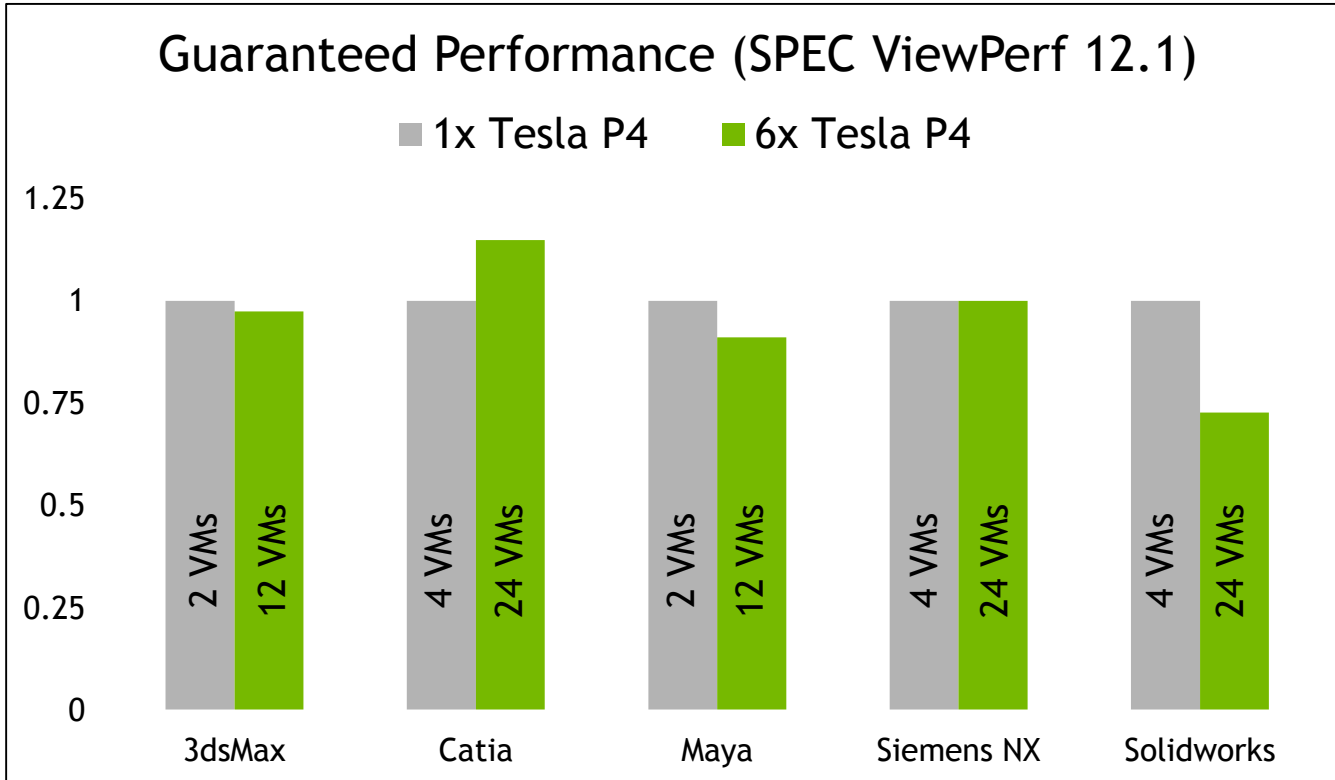


TESLA P4 BENEFITS
Performance
Price/Performance
Form Factor
Power Consumption
Pascal Benefits

* Tested on Dell R740 (2x Intel Xeon Gold 6154 CPU @ 3.0 GHz, 18 Cores and is based on geometric mean across 3dsMax, Catia, Maya, Siemens NX and Solidworks

New Intel CPU allows 6x Tesla P4

6x Users @ Comparable Guaranteed Performance



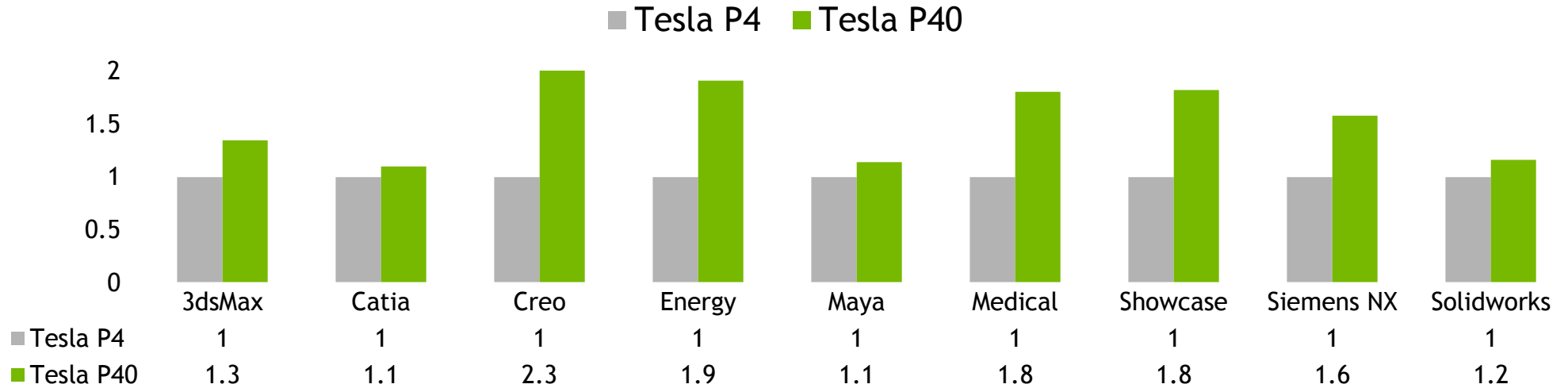
New Intel CPU (3GHz 18c)
allows the use of 6x P4s

Guaranteed performance is
close to the performance of
single P4

* Tested on a Dell R740 with 2x Intel Xeon Gold 6154 CPU @ 3.0 GHz, 18 Cores

P40 provides up to 2.3x more Perf than P4

SPEC ViewPerf 12.1 - Single VM (FRL-Off)



TESLA P4	TESLA P40
Many Low-Mid End Users	Few Mid-High End Users
Price/Performance	Performance
Form Factor	High Framebuffer Profiles (12GB and 24GB)
Power Consumption	
Multiple Profiles per Server (Many P4s)	

NVIDIA vGPU Scheduling Policies



Enterprise Customers

Best Effort Scheduler

default in Virtual GPU March 2018 Release (6.0)

Reason:

Maximum utilization of GPU cycles

Consider:

Equal Share Scheduler for
Compute Workloads
Delivering Guaranteed QoS



Cloud Service Providers

Fixed Share Scheduler

Reason:

Guaranteed QoS - Performance
GPU resources fenced off per profile

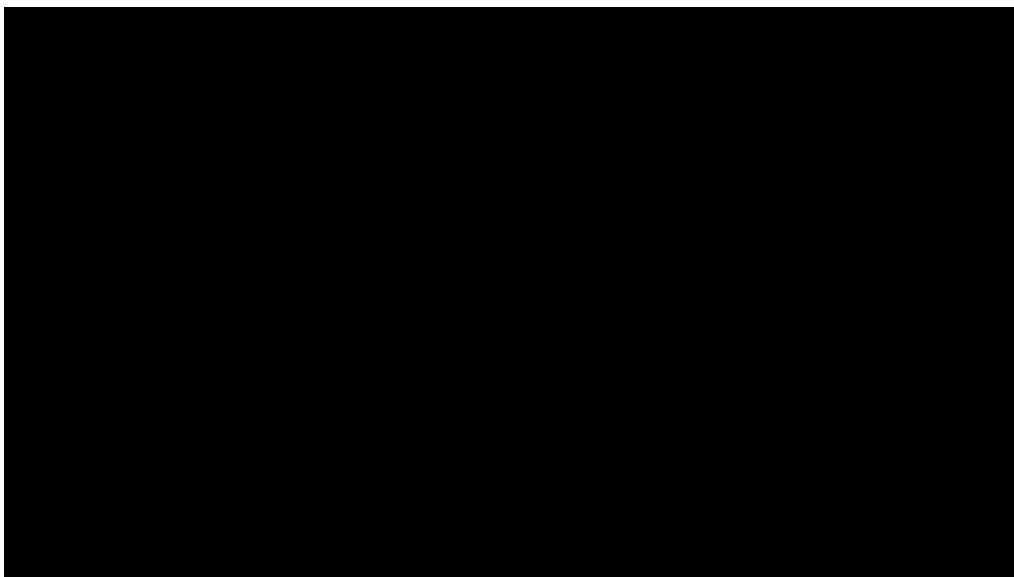
COMPARING THE SCHEDULING MODES

A high level summary cheat sheet

	BEST EFFORT	EQUAL SHARE	FIXED SHARE
Supported HW	Maxwell, Pascal	Pascal	Pascal
Primary Use cases	Enterprise	Enterprise	Cloud
vGPU aware	No	Yes	Yes
Needs mixed compute/graphics	Supported	Recommended	Recommended
Idle cycle redistribution	Yes	No	No
Guaranteed QoS	No	Yes	Yes
Noisy neighbor protection	No	Yes	Yes
FRL required	Yes	No	No

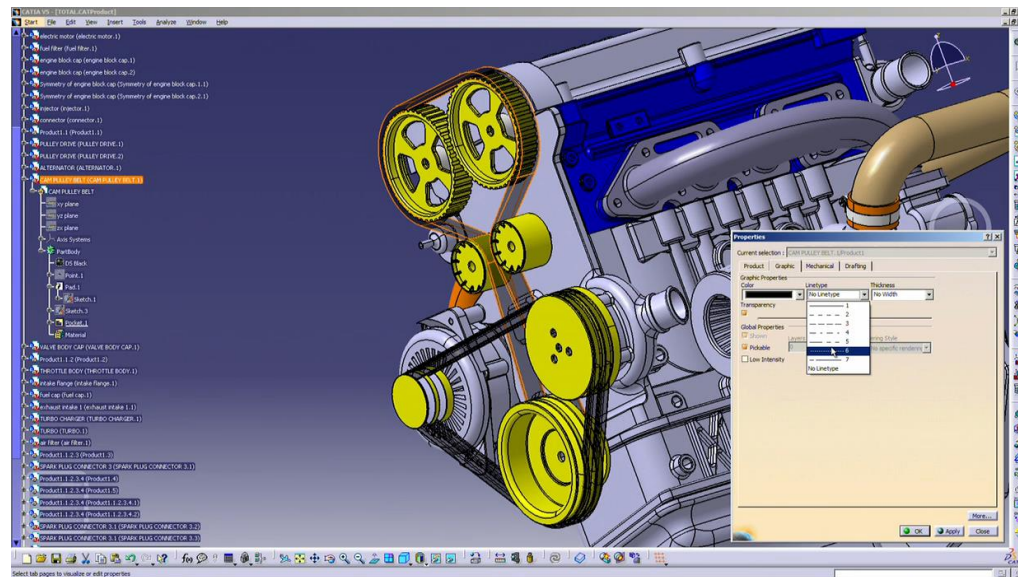
Benchmarking = Guaranteed Performance

Benchmark



Synthetic workload (4x Speed)

Human workflow



Human workflow (4x Speed)

Start with Guaranteed Performance ...

... explore individual scale for each customer during a POC

Defining Scale by Benchmarking

- Same Methodology as Quadro
- Familiar Methodology to the customer
- Guaranteed Performance
- Conservative Recommendation
- Allows Mapping Quadro boards

Defining Scale with real End Users

- Scale is individual to each customer
- Allows the Effect of time sharing
- Can lead to higher scale
- Performance at higher scale isn't guaranteed
- Leveraging the impact of time sharing requires Best-Effort Scheduling Policy

P1000 Class Catia Users (SPEC ViewPerf 12.1)*

Customer Experience**

4x Tesla P4

8 (4x 2)

~12-16 (4x 3-4)

* Tested on Dell R740 (2x Intel Xeon Gold 6154 CPU @ 3.0 GHz, 18 Cores)

GRID vPC

Defining User Experience (UX)

Remoted Frames

Describes the number of frames that are sent to the end user.

End-User Latency

Describes how remote the session feels or how interactive/laggy the session is.

Image Quality

Describes how much the image was impacted & manipulated by the remote protocol.

Functionality

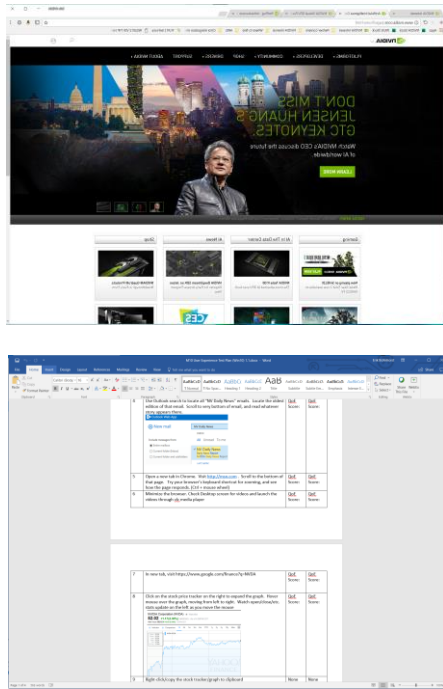
Describes if the remote desktop supports the same range of applications (API Support).

Consistency

Describes how much the user experience varies during the test run.

NVIDIA vPC Benchmark

Modern Apps

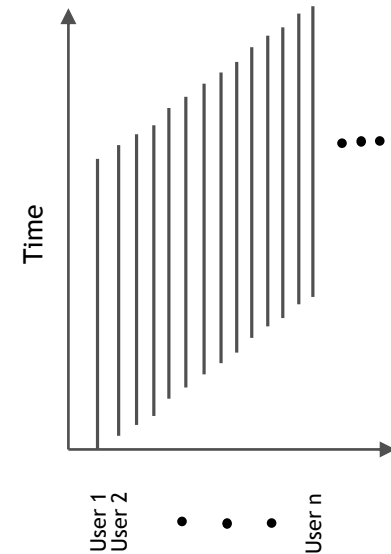


Many User, Many Behaviors

USER #1	USER #2
Google Chrome (Video)	MS Word 2016
Windows Media Player	Microsoft Edge (PDF)
MS Word 2016	MS Excel 2016
Microsoft Edge (PDF)	Google Chrome (Web)
MS Excel 2016	Google Chrome (Video)

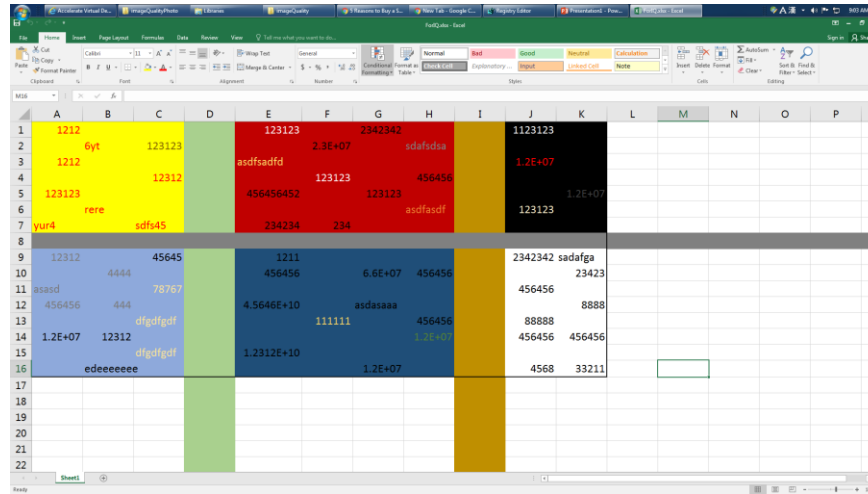
USER #3	USER #4	...
Windows Media Player	Google Chrome (Web)	
MS Word 2016	Google Chrome (Video)	
Microsoft Edge (PDF)	Windows Media Player	
MS Excel 2016	MS Word 2016	
Google Chrome (Web)	Microsoft Edge (PDF)	

Different Timing



Horizon 7 Image Quality Improvements

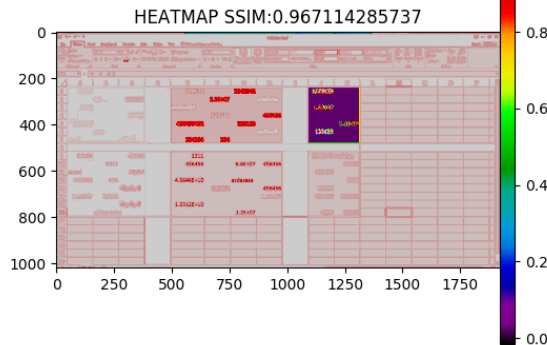
Reference
image



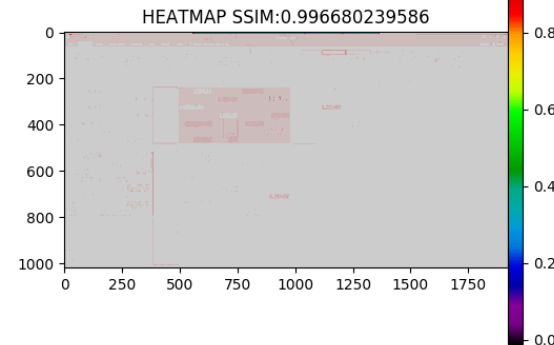
The screenshot shows an Excel spreadsheet with a grid of data. The grid is divided into several colored regions: yellow (top-left), green (top-middle), red (top-right), blue (middle-left), and dark blue (middle-right). The data includes alphanumeric strings and numerical values in scientific notation. The spreadsheet interface includes the ribbon, formula bar, and status bar.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	1212				123123		2342342			1123123						
2	6yt	123123				2.3E+07	sdfasda									
3	1212				asdfsdf		123123	456456		1.2E+07						
4	123123		12312		456456452		123123				1.2E+07					
5		rere				234234	234	asdfsdf			123123					
6	yu4		sdfs45													
7																
8	12312		45645		1211					2342342	sadafga					
9		4444			456456		6.6E+07	456456				23423				
10	asasd		78767							456456		8888				
11	456456	444	djgdjgd		4.5646E+10		asdasaaa	456456				88888				
12							1111111									
13	1.2E+07	12312	djgdjgd					456456				456456	456456			
14					1.2312E+10			1.2E+07								
15			edeeeeeee							4568	33211					
16																
17																
18																
19																
20																
21																
22																

YUV 4:2:0

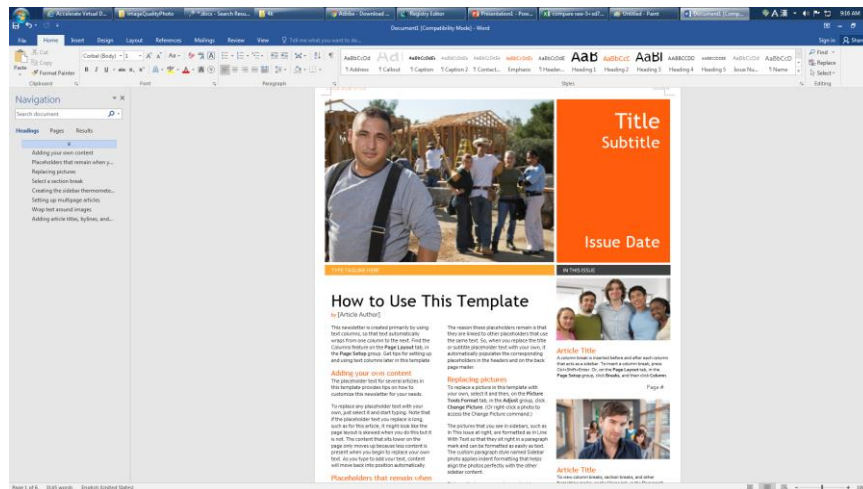


YUV 4:4:4



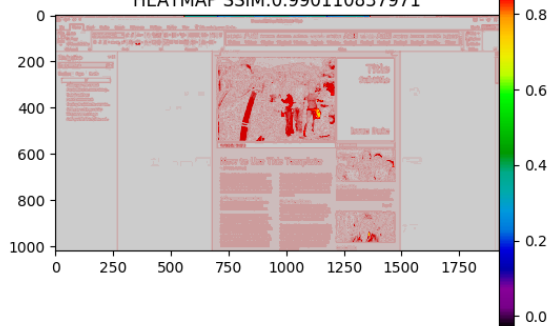
Horizon 7 Image Quality Improvements

Reference image



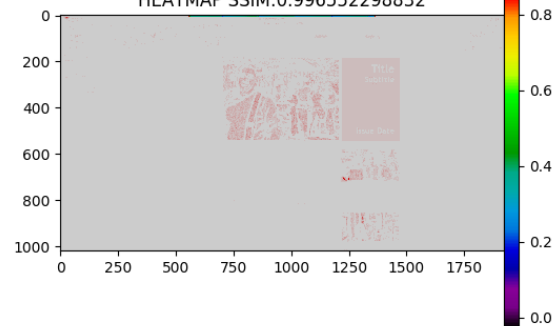
YUV 4:2:0

HEATMAP SSIM:0.990110837971



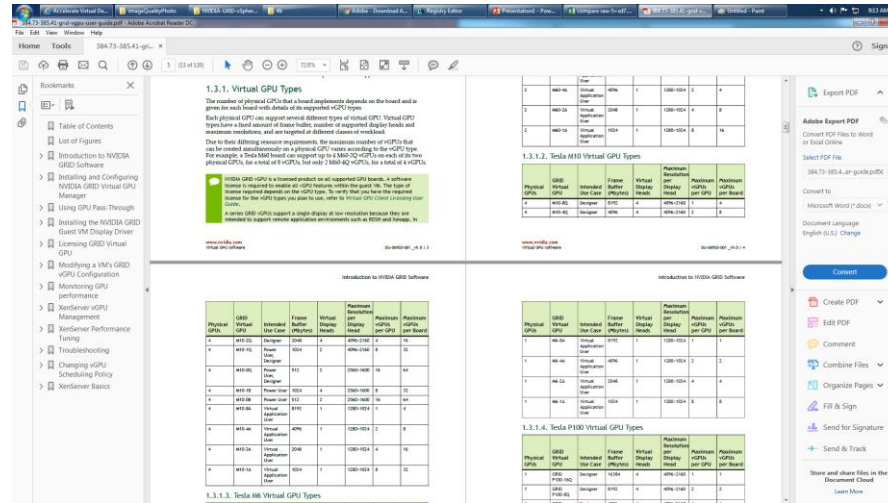
YUV 4:4:4

HEATMAP SSIM:0.996552298832

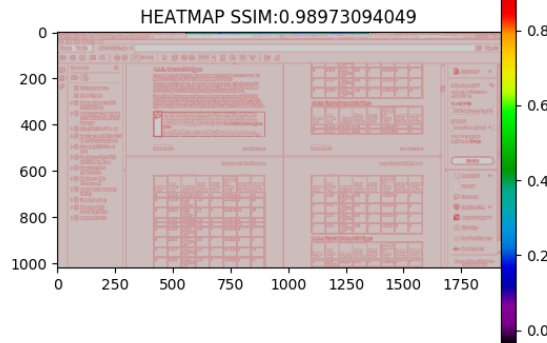


Horizon 7 Image Quality Improvements

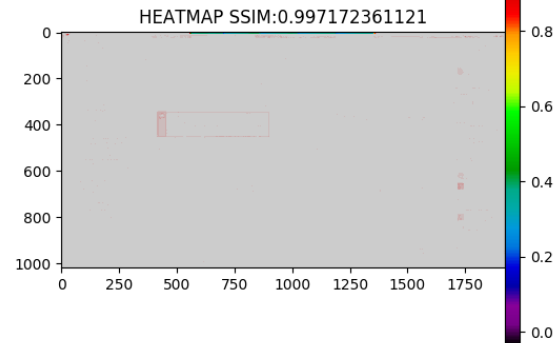
Reference image



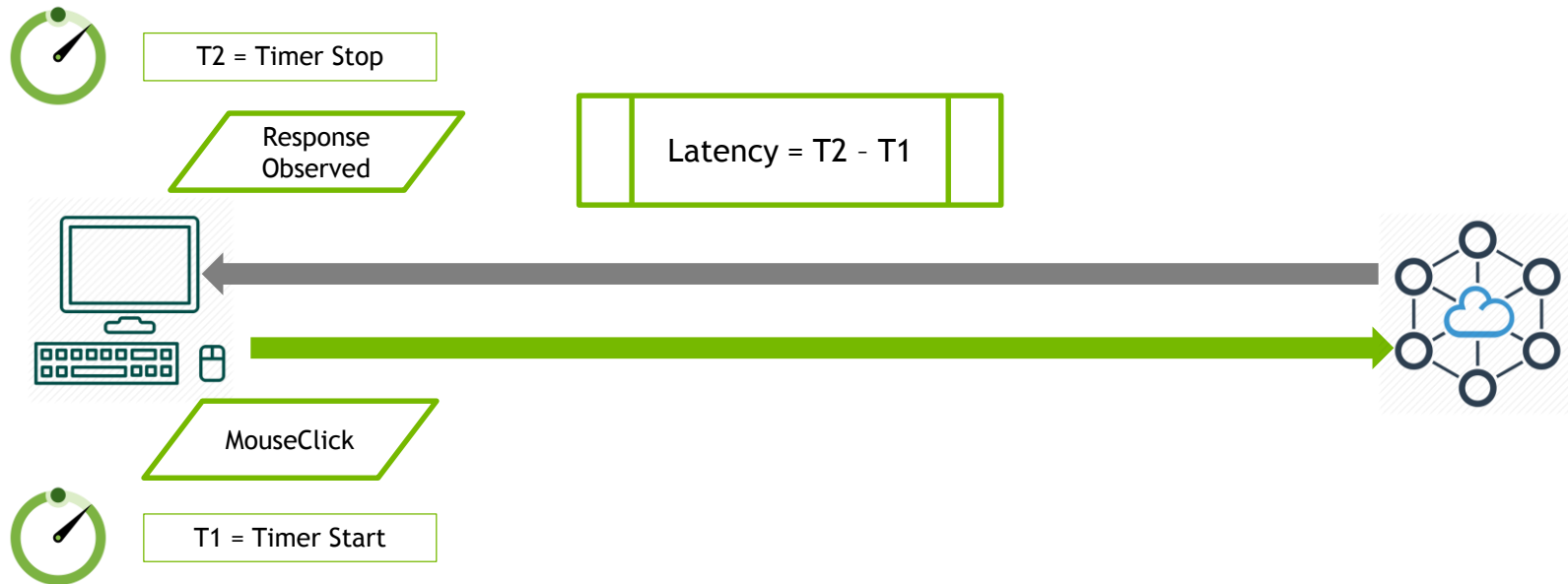
YUV 4:2:0



YUV 4:4:4



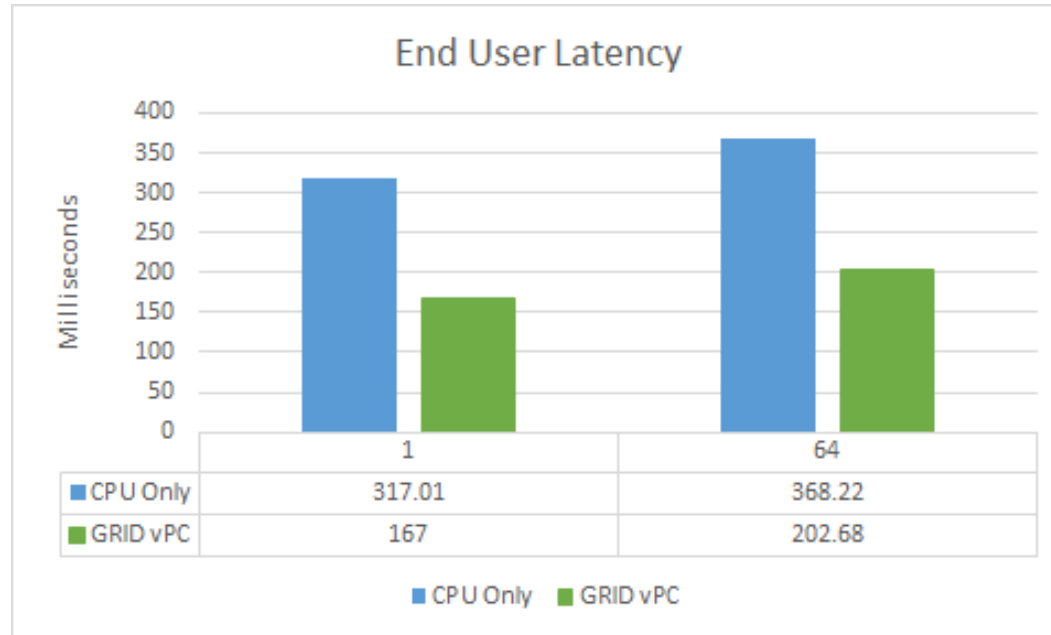
End User Latency (Click-To-Photon)



Best End-User Latency with NVIDIA vPC

Decrease of 140-160ms for best remoted user experience

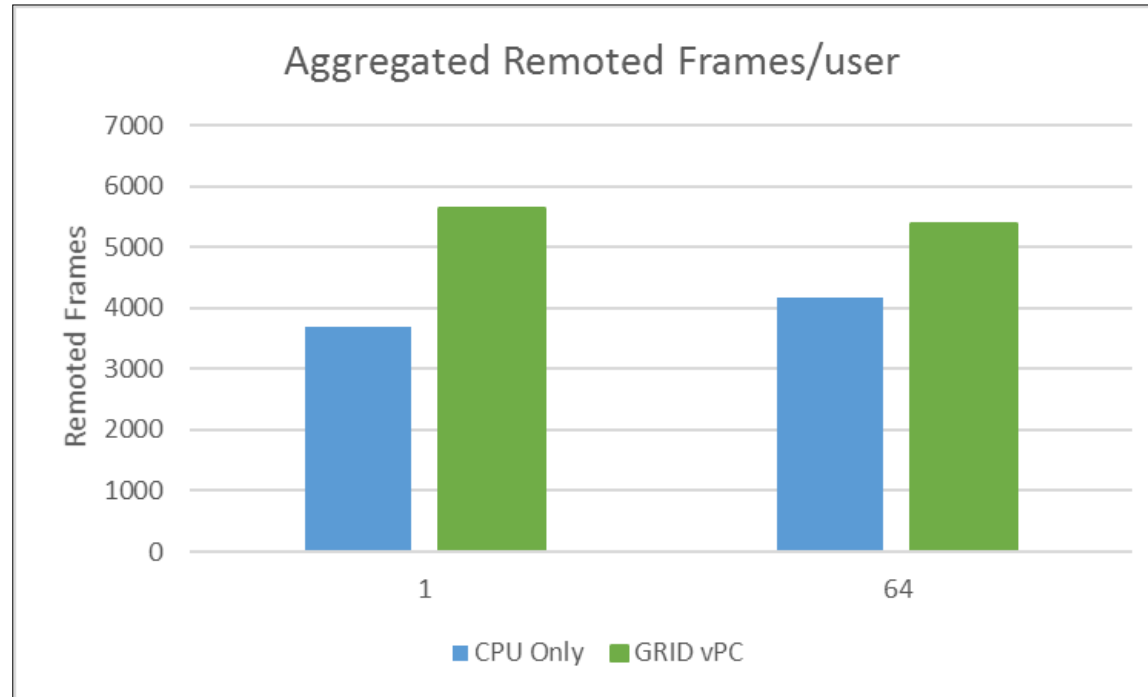
VMware Horizon 7.4 (YUV 4:4:4)



- End-User Latency decrease of 140ms with 1VM
- End-User Latency decrease of 160ms with 64 VMs

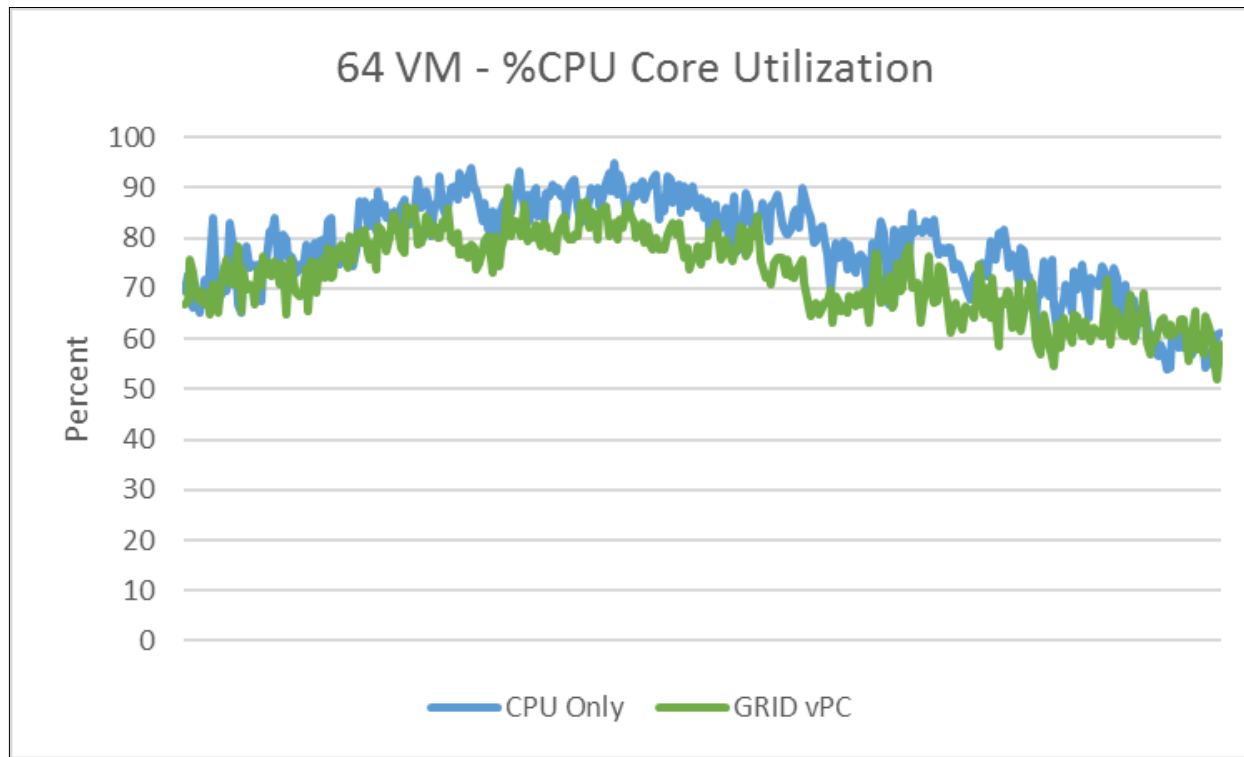
40% More Remoted Frames with GRID vPC

VMware Horizon 7.4 (YUV 4:4:4)



Up to 25% CPU offload for Highest Density

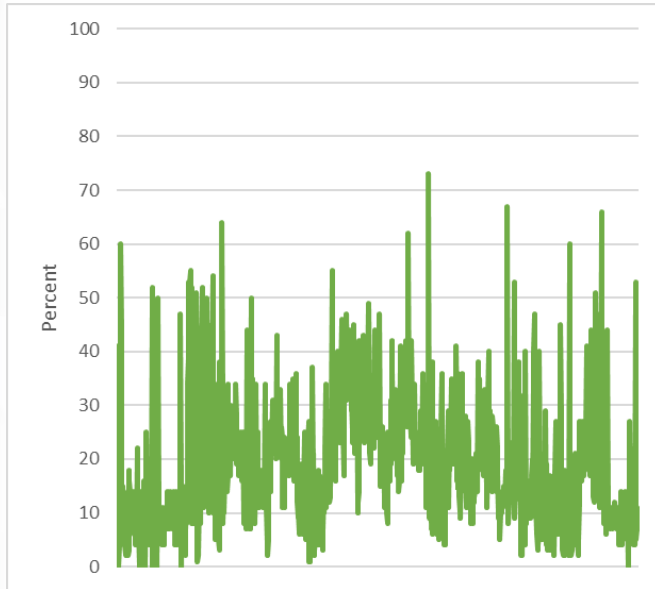
VMware Horizon 7.4 (YUV 4:4:4)



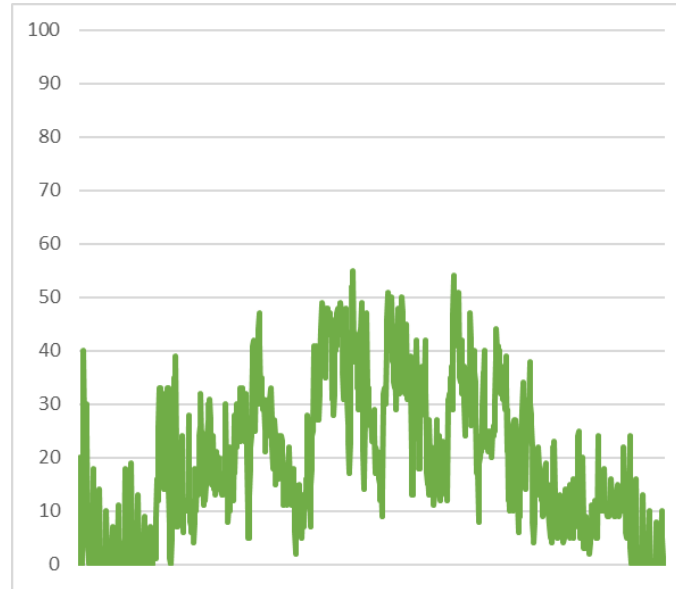
TESLA M10 MEETS THE NEEDS OF KNOWLEDGE WORKERS

Tesla M10 GPU and Encode Engine match the needs of Windows 10

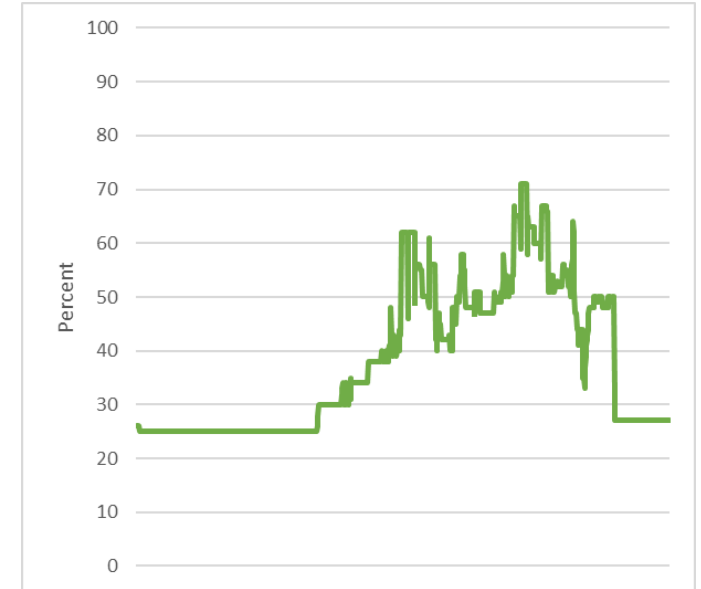
Tesla M10 GPU Utilization for 32 VMs (8/GPU)



Tesla M10 Encoder Utilization for 32 VMs (8/GPU)



VM Framebuffer Utilization M10-1B



Cirrus Knowledge Worker Workload (Excel, Word, PowerPoint, Chrome, Media Player, PDF) with VMware Horizon 7.4 YUV 4:4:4

NVIDIA GRID VGPU FOR HIGHEST DENSITY AND BEST USER EXPERIENCE

Best User
Experience

Highest Density

Tesla M10 for
Win10



THANK YOU