S9881 Using Industry Standard Benchmark Tools to Size Graphics Accelerated Applications

Mike Brennan, Product Manager, Virtual Client Computing and Graphics
Vadim Lebedev, Technical Marketing Engineer
21 March 2019

What we will cover

- Let’s talk about the basics on virtualizing pro graphics apps
- How do you measure performance?
- Key NVIDIA cards
- Sample benchmark performance
- Server/GPU performance
- Where do I start with sizing?
- Cisco lineup
- Key takeaways
- Q&A

Lets talk basics

Can you virtualize Catia and SolidWorks?

Yes you can!

Cisco has 14 hardware/ software combinations certified

[Dassault VDI Certifications]()

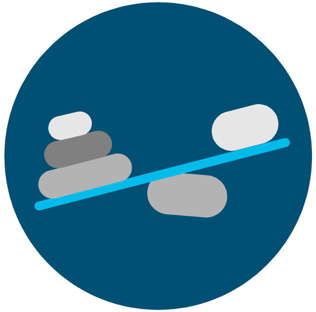
CATIA

SOLIDWORKS

DASSAULT SYSTEMES

The 3DEXPERIENCE® Company

Why replace physical graphics workstations?



Key performance requirements for Virtual Workstations

- User requirements
- Software requirements
- Display resolution
- Monitors
- CPU and memory performance
- Graphics card oversubscription
- Frame rate
- Multi-user graphics card scheduling engine

User requirements

- User role
- Concurrent applications open
- Complexity of graphic application
- Collaboration requirements
- Working hours

User roles

Light user type

- Primarily read only – documentation, project managers
- Small subsets of entire entity

Medium user type

- Read only and design
- Small and medium sub-assemblies

Heavy user type

- Design and render
- Large sub-assemblies and full model

Software requirements

- Dassault minimum requirements
- Dassault support for virtualization
- Dassault hardware qualification
- Dassault delivery partners
- Dassault support model

Display resolution and monitors

- Maximum display resolution supported
- Display resolution expected by user type
- Number of monitors per user supported
- Number of monitors expected by user type

CPU and memory performance

CPU selection criteria

- Total frequency in MHz/CPU and server
- Core count
- Planned user count

Memory selection criteria

- Balance
- Frequency

Graphics card oversubscription

User count per graphics card

- Fixed at GPU frame buffer divided by vGPU profile
  - For an NVIDIA P4 card
  - For a 2Q profile: 8GB frame buffer/2GB frame buffer per user = 4 Users per card.

GPU oversubscription

- NVIDIA concept
- Based on scheduler chosen
- For the T4 card, light user could get more than 12.5% of GPU resources

Frame rates

The great equalizer for performance

- For computer video displays,
  - frame rate = #frames or images displayed per second
- For a given application
  - Provides a mechanism to compare systems performance
  - Describes a mechanism by which system requirements can be stated

Virtual Graphics Workstation insights

- Frame rate can be controlled – or not
- Frame rate can be set in the NVIDIA and Desktop Broker software policy

Multi-user graphics cards scheduling engines

NVIDIA supports 3 models

Best effort (default)

- User gets GPU resources based on current availability
- At any given point in time a user MIGHT get more than his fair share of GPU

Fixed share

- Each user gets the same dedicated performance at all times

Equal share

- Each VM gets and equal share of the GPU resources

How do you measure performance?

Performance measurement

Industry graphic benchmark examples

- SPECviewperf 13
- PassMark Software
- Unigen Heaven, Valley, etc
- Others

SPECviewperf 13 supports nine Virtual Professional Graphics Applications

SPECviewperf 13 provides a composite benchmark score across all nine applications

SPECviewperf 13 provides capability to score individual applications

SPECviewperf 13 provides ability to measure performance across various graphic card, CPU, memory, scheduling and frame rate scenarios

SPECviewperf 13 has the following minimum requirements:

- ✓ Microsoft Windows 10 64-bit RS3 or later VM
- ✓ OpenGL 4.0
- ✓ Direct X12 support
- ✓ 8GB of installed system memory
- ✓ 80GB available disk space
- ✓ 1920x1080 screen resolution for submissions published on the SPEC website

Applications driving large TAM, Verticals

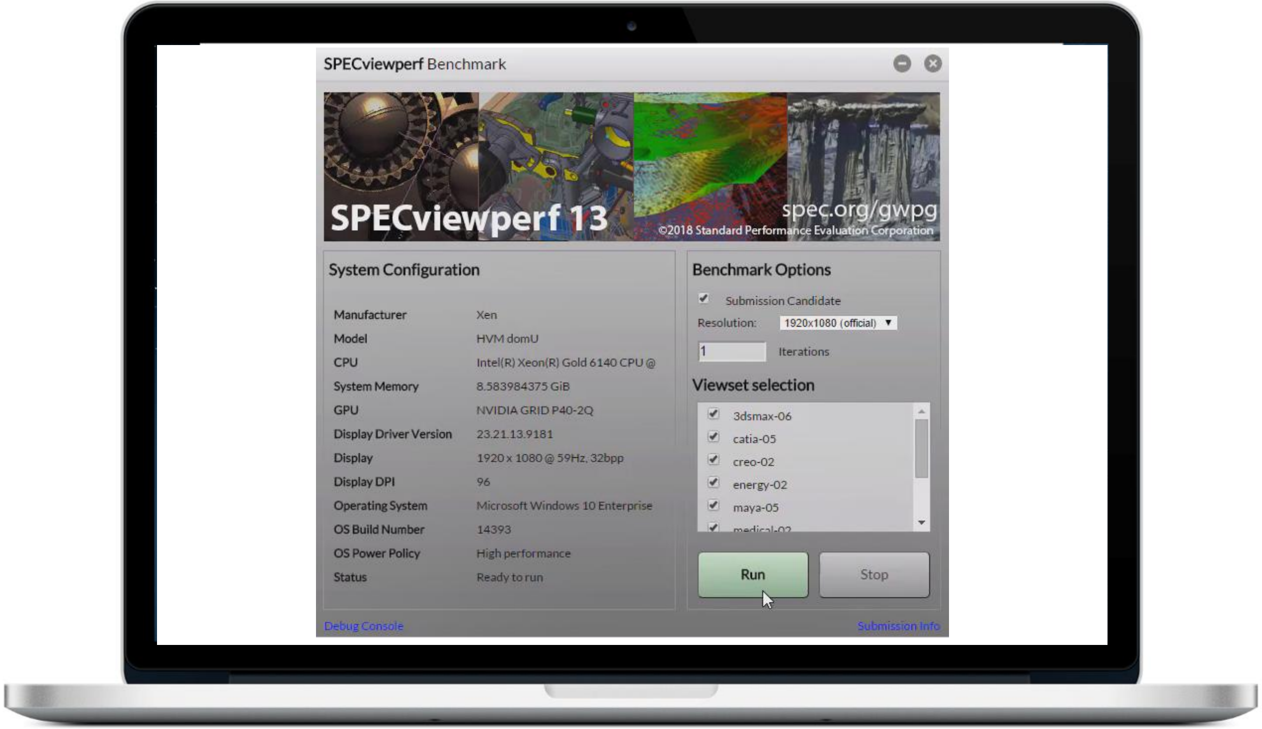
2d/3d design, video	Engineering	Medical	Oil exploration
MAYA	PLM Software SIEMENS NX	SCI	Energy (OpendTect)
3 MAX	SOLIDWORKS CATIA	ImageVis3D	dGB Earth Sciences
SHOWCASE	creo®		

Manufacturing, Auto

Oil & Gas

Utilities, Sled, Arch/Design/Constr

SPECviewperf 13 Test Console



System Configuration

Manufacturer	Xen
Model	HVM domU
CPU	Intel(R) Xeon(R) Gold 6140 CPU @
System Memory	8.583984375 GiB
GPU	NVIDIA GRID P40-2Q
Display Driver Version	23.21.13.9181
Display	1920 x 1080 @ 59Hz, 32bpp
Display DPI	96
Operating System	Microsoft Windows 10 Enterprise
OS Build Number	14393
OS Power Policy	High performance
Status	Ready to run

Benchmark Options

☑ Submission Candidate

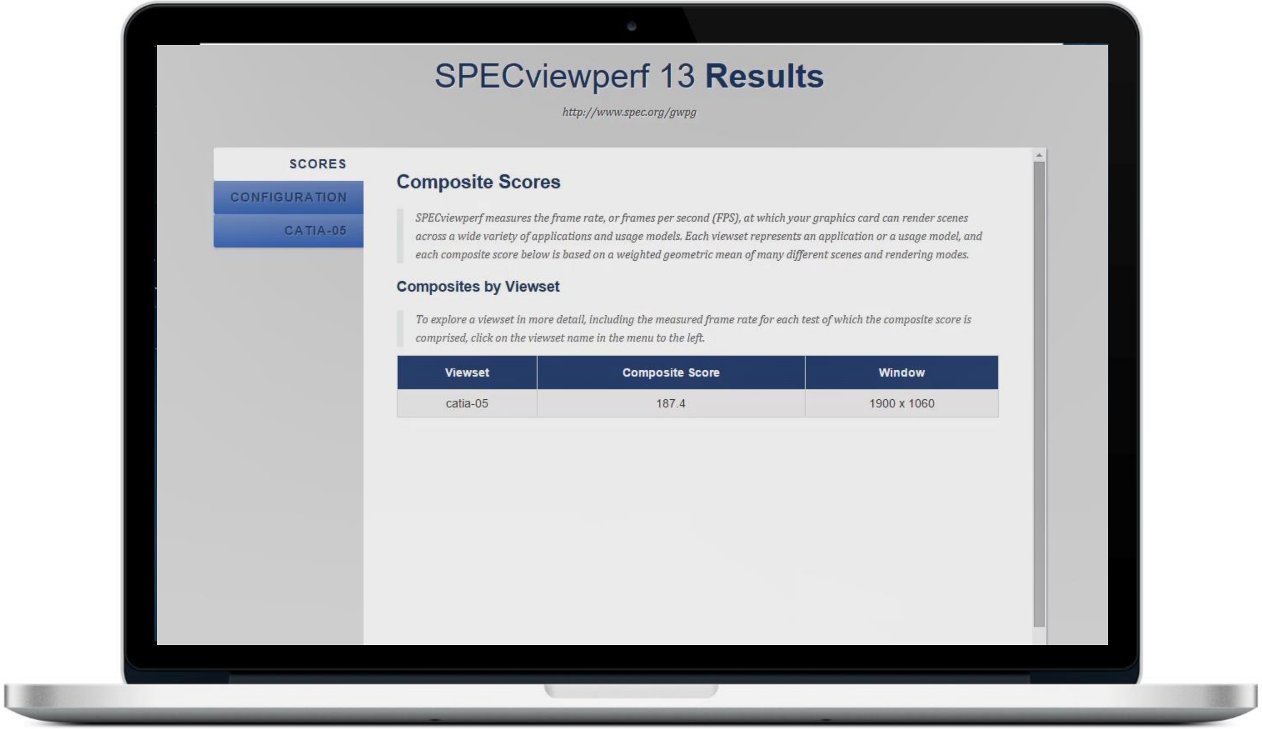
Resolution: 1920x1080 (official)

1 Iterations

Viewset selection

☑ 3dsmax-06
☑ catia-05
☑ creo-02
☑ energy-02
☑ maya-05
☑ medical-02

SPECviewperf 13 Test Results - Composite



SPECviewperf 13 **Results**

http://www.spec.org/gwpg

Composite Scores

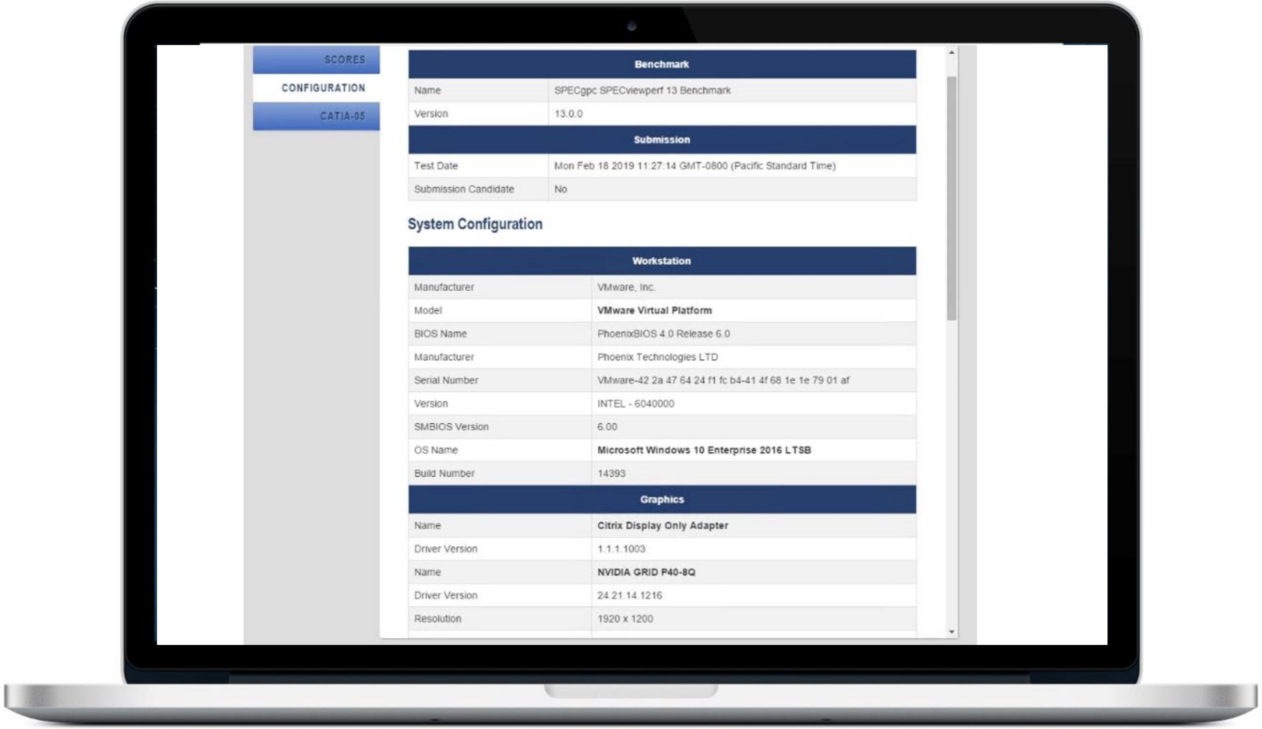
> *SPECviewperf measures the frame rate, or frames per second (FPS), at which your graphics card can render scenes across a wide variety of applications and usage models. Each viewset represents an application or a usage model, and each composite score below is based on a weighted geometric mean of many different scenes and rendering modes.*

Composites by Viewset

> *To explore a viewset in more detail, including the measured frame rate for each test of which the composite score is comprised, click on the viewset name in the menu to the left.*

Viewset	Composite Score	Window
catia-05	187.4	1900 x 1060

SPECviewperf 13 Test Results - Configuration



Benchmark	
Name	SPECgpc SPECviewperf 13 Benchmark
Version	13.0.0

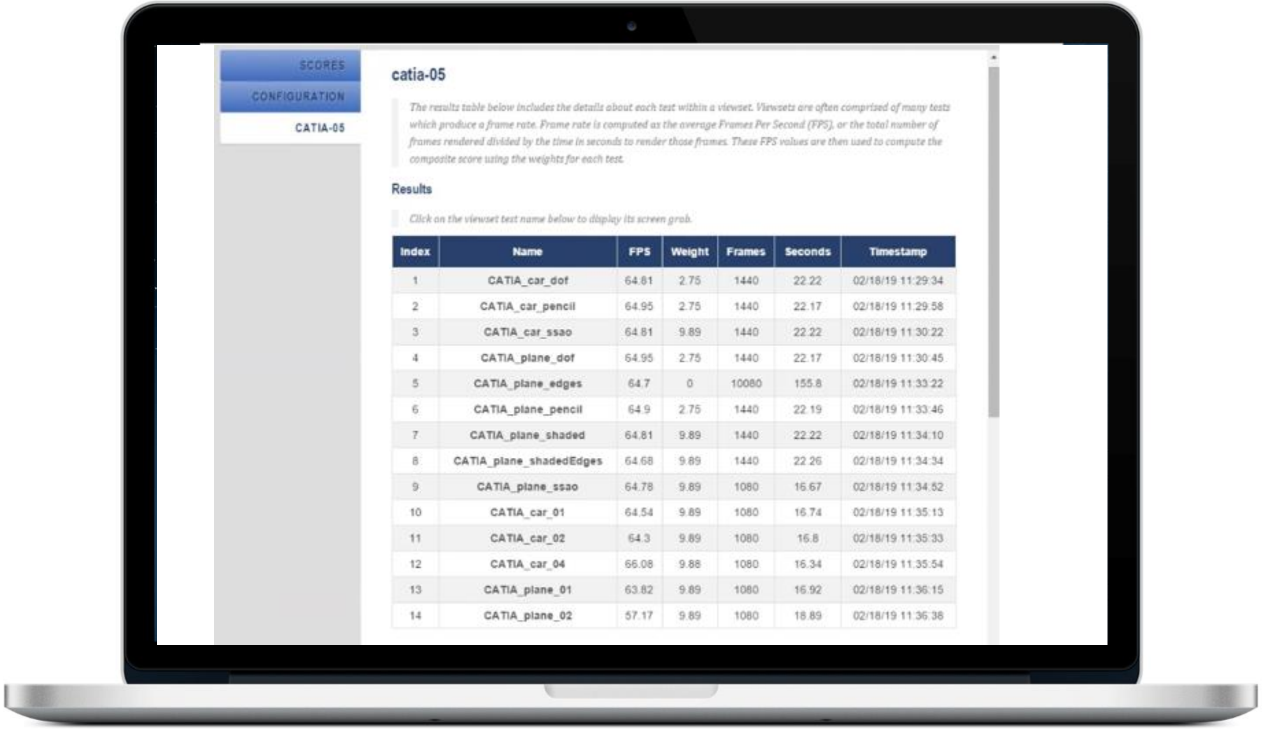
Submission	
Test Date	Mon Feb 18 2019 11:27:14 GMT-0800 (Pacific Standard Time)
Submission Candidate	No

System Configuration

Workstation	
Manufacturer	VMware, Inc.
Model	**VMware Virtual Platform**
BIOS Name	PhoenixBIOS 4.0 Release 6.0
Manufacturer	Phoenix Technologies LTD
Serial Number	VMware-42 2a 47 64 24 f1 fc b4-41 4f 68 1e 1e 79 01 af
Version	INTEL - 6040000
SMBIOS Version	6.00
OS Name	**Microsoft Windows 10 Enterprise 2016 LTSB**
Build Number	14393

Graphics	
Name	**Citrix Display Only Adapter**
Driver Version	1.1.1.1003
Name	**NVIDIA GRID P40-8Q**
Driver Version	24.21.14.1216
Resolution	1920 x 1200

SPECviewperf 13 Test Results – Viewset Catia



catia-05

> *The results table below includes the details about each test within a viewset. Viewsets are often comprised of many tests which produce a frame rate. Frame rate is computed as the average Frames Per Second (FPS), or the total number of frames rendered divided by the time in seconds to render those frames. These FPS values are then used to compute the composite score using the weights for each test.*

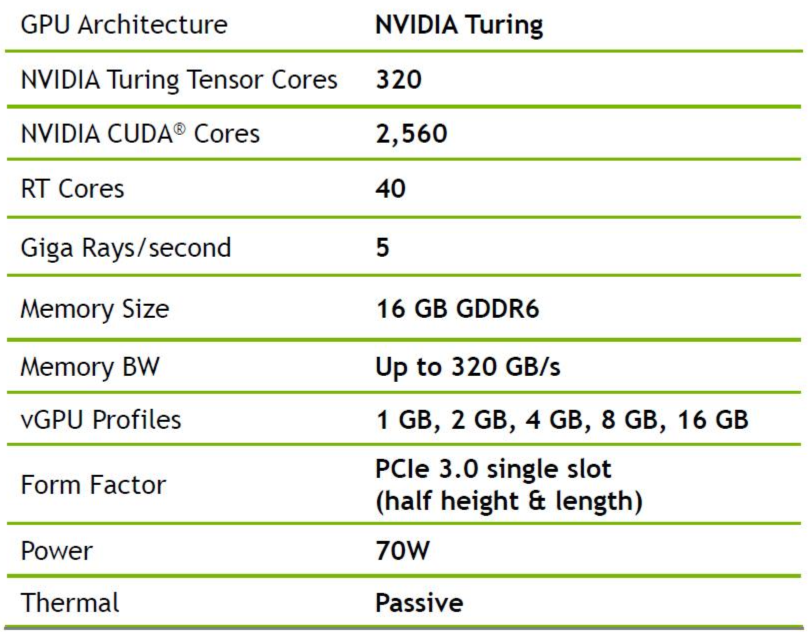
Results

> *Click on the viewset test name below to display its screen grab.*

Index	Name	FPS	Weight	Frames	Seconds	Timestamp
1	CATIA_car_dof	64.81	2.75	1440	22.22	02/18/19 11:29:34
2	CATIA_car_pencil	64.95	2.75	1440	22.17	02/18/19 11:29:58
3	CATIA_car_ssao	64.81	9.89	1440	22.22	02/18/19 11:30:22
4	CATIA_plane_dof	64.95	2.75	1440	22.17	02/18/19 11:30:45
5	CATIA_plane_edges	64.7	0	10080	155.8	02/18/19 11:33:22
6	CATIA_plane_pencil	64.9	2.75	1440	22.19	02/18/19 11:33:46
7	CATIA_plane_shaded	64.81	9.89	1440	22.22	02/18/19 11:34:10
8	CATIA_plane_shadedEdges	64.68	9.89	1440	22.26	02/18/19 11:34:34
9	CATIA_plane_ssao	64.78	9.89	1080	16.67	02/18/19 11:34:52
10	CATIA_car_01	64.54	9.89	1080	16.74	02/18/19 11:35:13
11	CATIA_car_02	64.3	9.89	1080	16.8	02/18/19 11:35:33
12	CATIA_car_04	66.08	9.88	1080	16.34	02/18/19 11:35:54
13	CATIA_plane_01	63.82	9.89	1080	16.92	02/18/19 11:36:15
14	CATIA_plane_02	57.17	9.89	1080	18.89	02/18/19 11:36:38

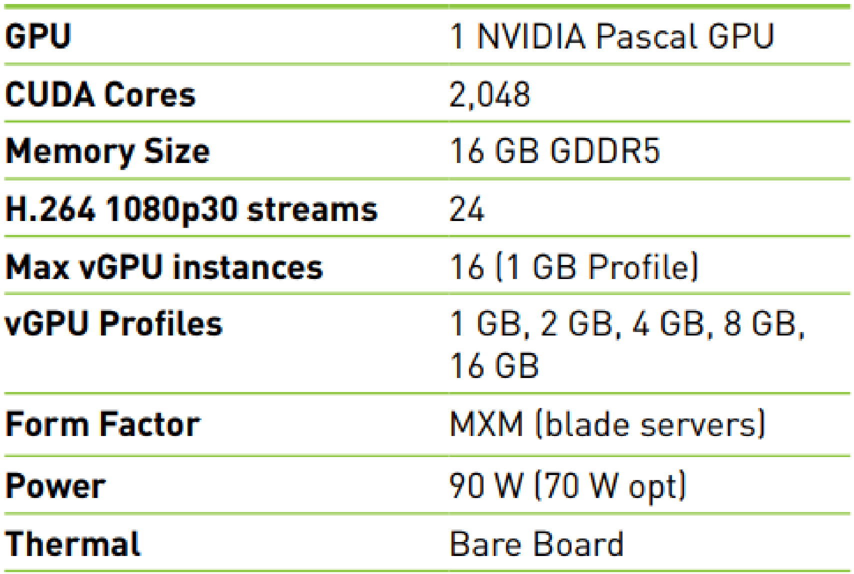
NVIDIA Tesla T4 and P40

Tesla T4 Key Specifications



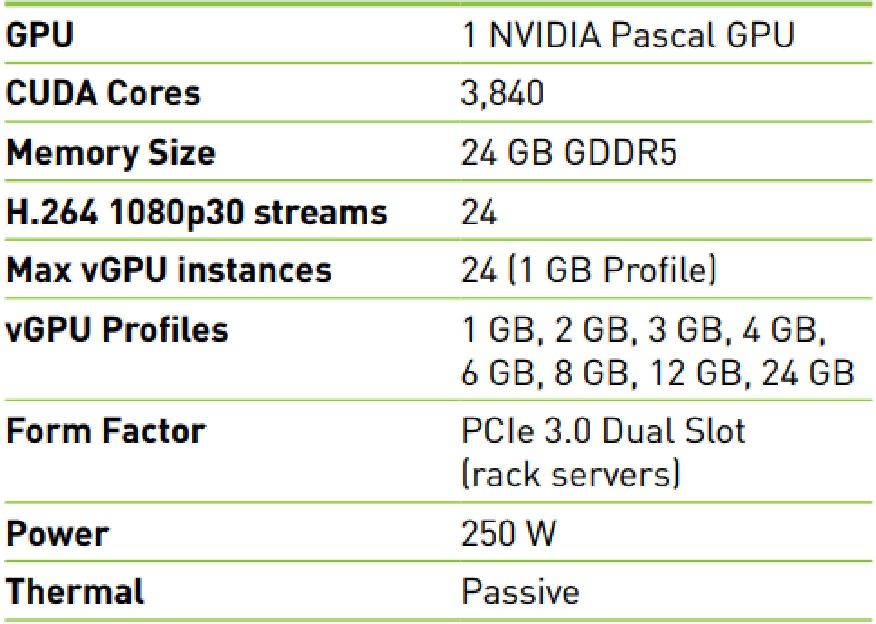
GPU Architecture	**NVIDIA Turing**
NVIDIA Turing Tensor Cores	**320**
NVIDIA CUDA® Cores	**2,560**
RT Cores	**40**
Giga Rays/second	**5**
Memory Size	**16 GB GDDR6**
Memory BW	**Up to 320 GB/s**
vGPU Profiles	**1 GB, 2 GB, 4 GB, 8 GB, 16 GB**
Form Factor	**PCIe 3.0 single slot (half height & length)**
Power	**70W**
Thermal	**Passive**

Tesla P6 Key Specifications



GPU	1 NVIDIA Pascal GPU
CUDA Cores	2,048
Memory Size	16 GB GDDR5
H.264 1080p30 streams	24
Max vGPU instances	16 (1 GB Profile)
vGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB
Form Factor	MXM (blade servers)
Power	90 W (70 W opt)
Thermal	Bare Board

Tesla P40 Key Specifications



GPU	1 NVIDIA Pascal GPU
CUDA Cores	3,840
Memory Size	24 GB GDDR5
H.264 1080p30 streams	24
Max vGPU instances	24 (1 GB Profile)
vGPU Profiles	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB
Form Factor	PCIe 3.0 Dual Slot (rack servers)
Power	250 W
Thermal	Passive

NVIDIA TESLA GPUs

Recommended for Virtualization

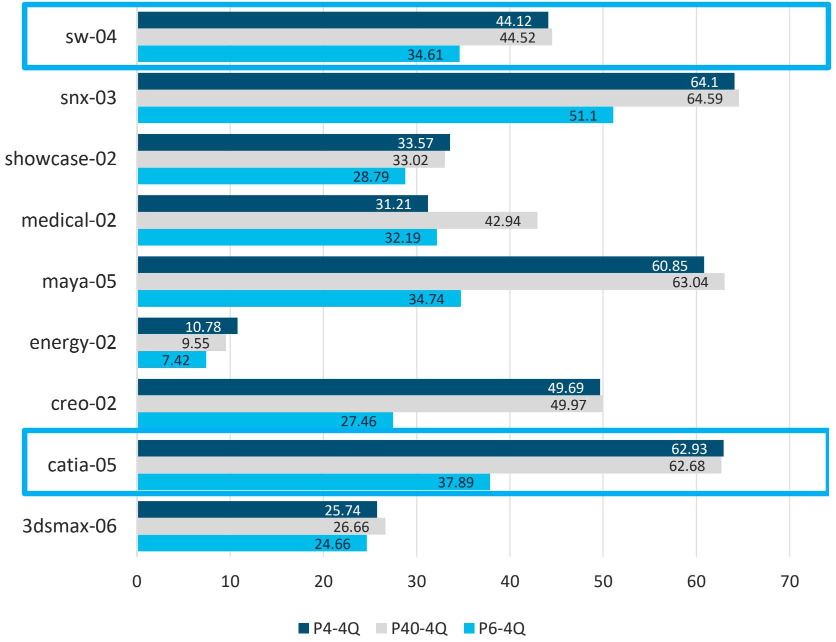
	V100	P40	T4	P4	M60	M10	P6
GPUs / Board (Architecture)	1 (Volta)	1 (Pascal)	1 (Turing)	1 (Pascal)	2 (Maxwell)	4 (Maxwell)	1 (Pascal)
CUDA Cores	5,120	3,840	2,560	2,560	4,096 (2,048 per GPU)	2,560 (640 per GPU)	2,048
Memory Size	32 GB/16 GB HBM2	24 GB GDDR5	16 GB GDDR6	8 GB GDDR5	16 GB GDDR5 (8 GB per GPU)	32 GB GDDR5 (8 GB per GPU)	16 GB GDDR5
vGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB, 32 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB	1 GB, 2 GB, 4 GB, 8 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB
Form Factor	PCIe 3.0 Dual Slot & SXM2 (rack servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	MXM (blade servers)
Power	250W/300W	250W	70W	75W	300W (225W opt)	225W	90W
Thermal	passive	passive	passive	passive	active/passive	passive	bare board
	PERFORMANCE Optimized					**DENSITY** Optimized	**BLADE** Optimized

Example Benchmark Insights

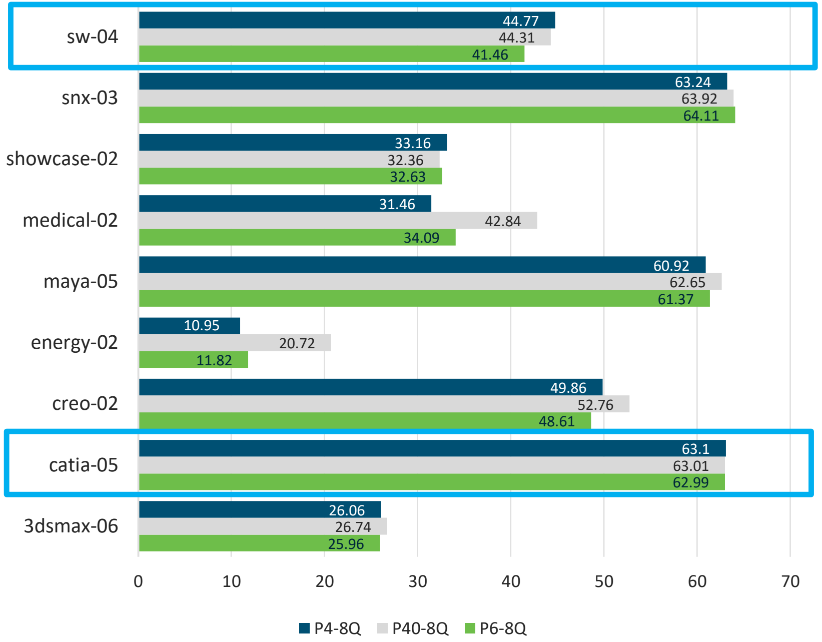
SPECviewperf 13 benchmark results

Compare three cards, two profiles, 1 VM, best effort, FRL On, Xeon 6140

1 VM 4Q on XenServer Host

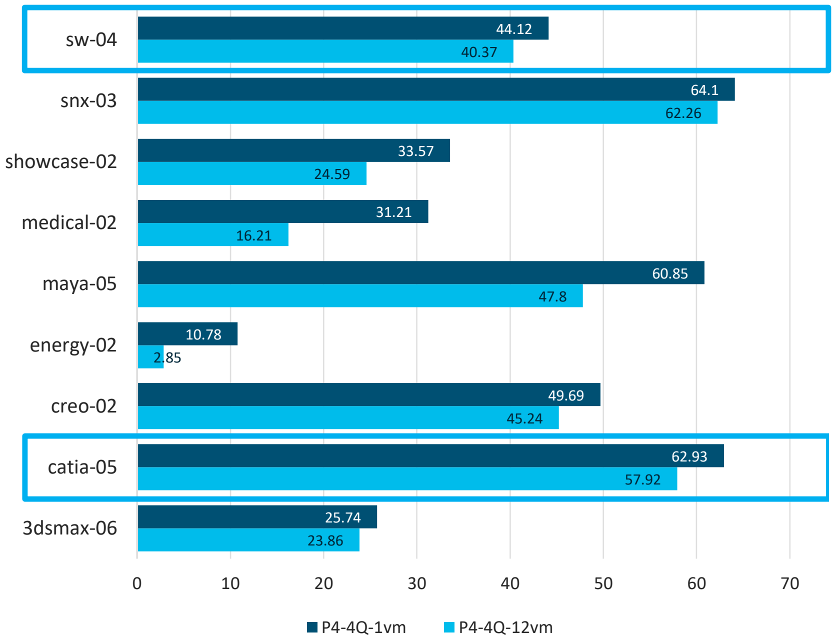


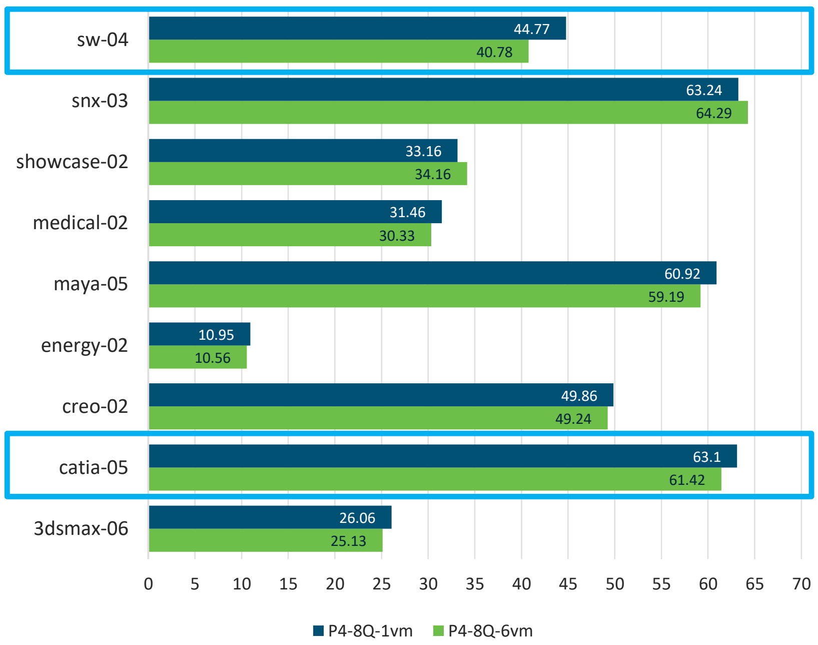
1 VM 8Q on XenServer Host



SPECviewperf 13 benchmark results

Compare one card, two profiles, 1 VM & max VMs, best effort and FRL On, Xeon 6140

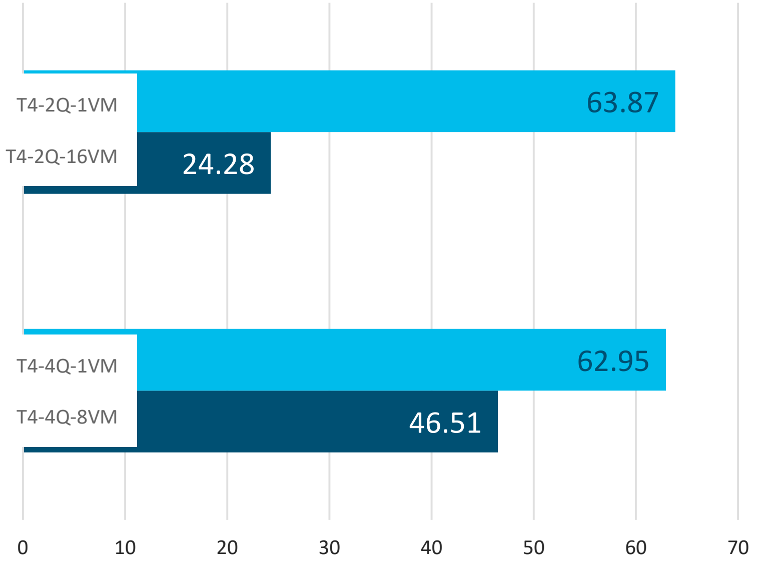

P-4 4Q profile testing


P-4 8Q profile testing

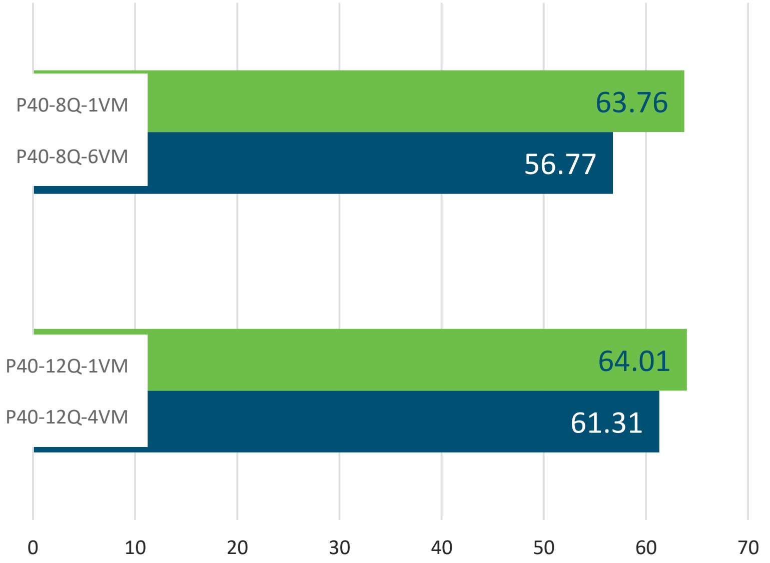
SPECviewperf 13 non-benchmark results - Catia

Compare two cards, 1 VM & max VMs, best effort and FRL On, Xeon 6140

T-4 2Q, 4Q profile testing



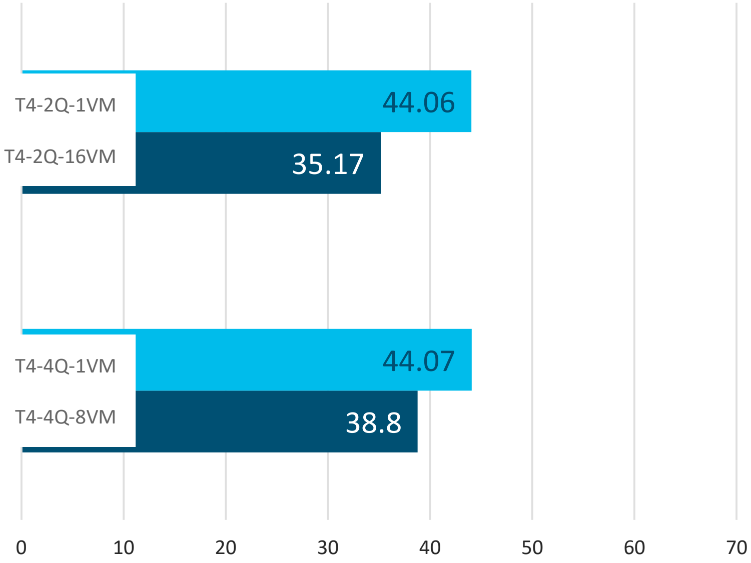
P40 8Q, 12Q



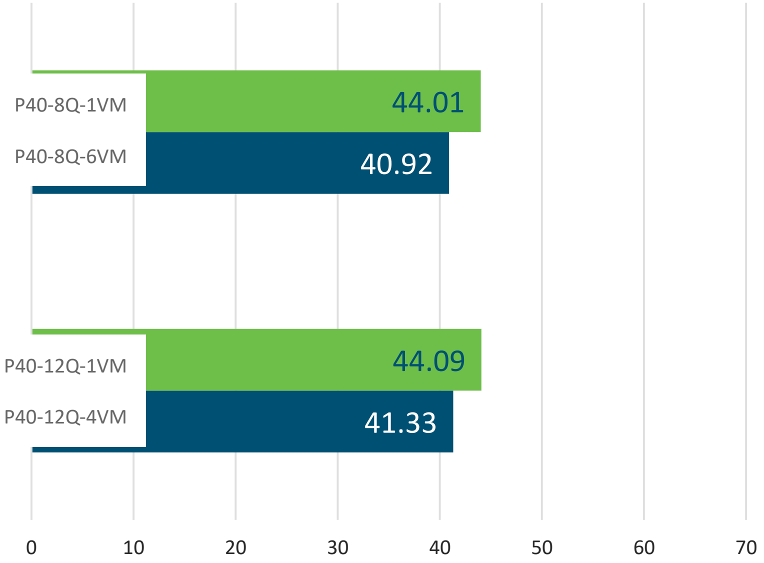
SPECviewperf 13 non-benchmark results -Solidworks

Compare two cards, 1 VM & max VMs, best effort and FRL On, Xeon 6140

T-4 2Q, 4Q profile testing



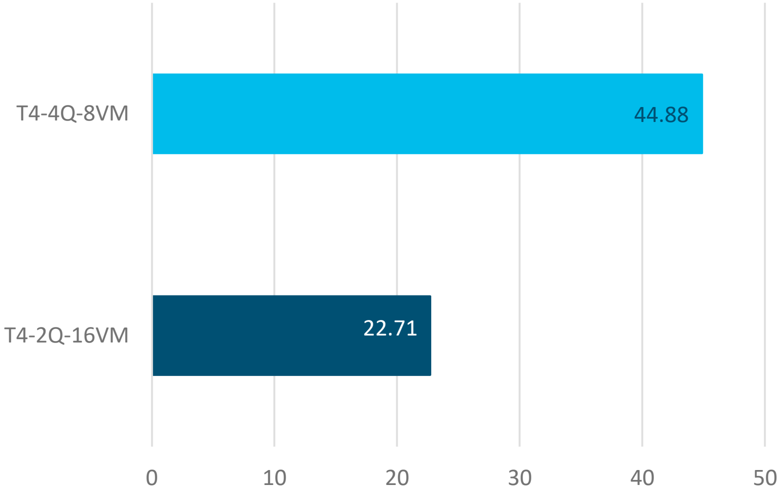
P40 8Q, 12Q



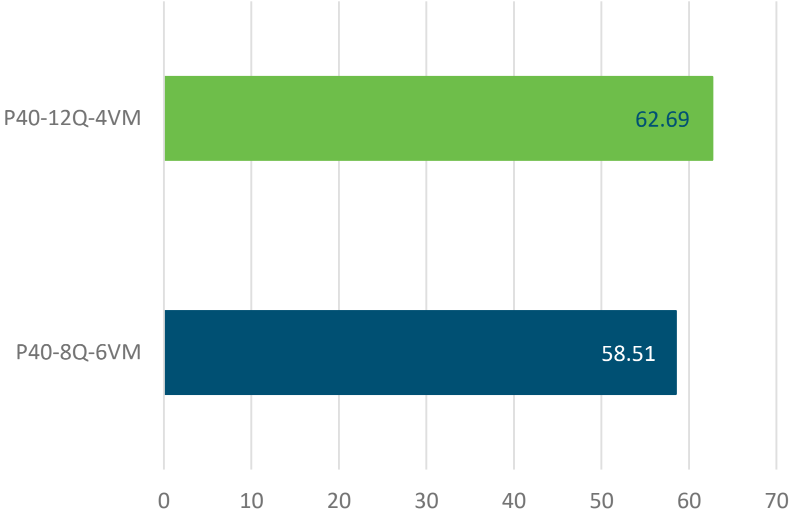
SPECviewperf 13 non-benchmark results - Catia

Compare two cards, max VMs, best effort and FRL On, Xeon 6136, 6128

T-4 2Q, 4Q profile, XEON 6136



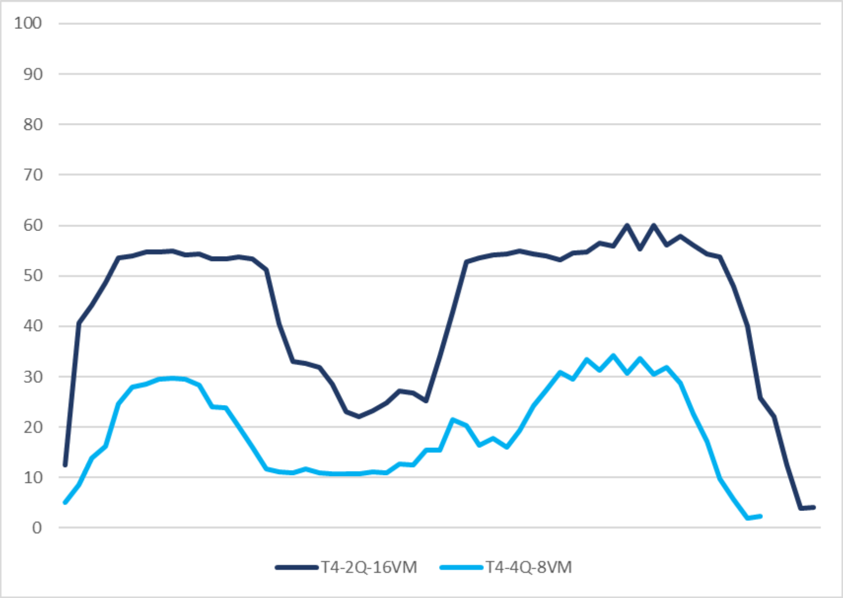
P40 8Q, 12Q Profile, XEON 6128



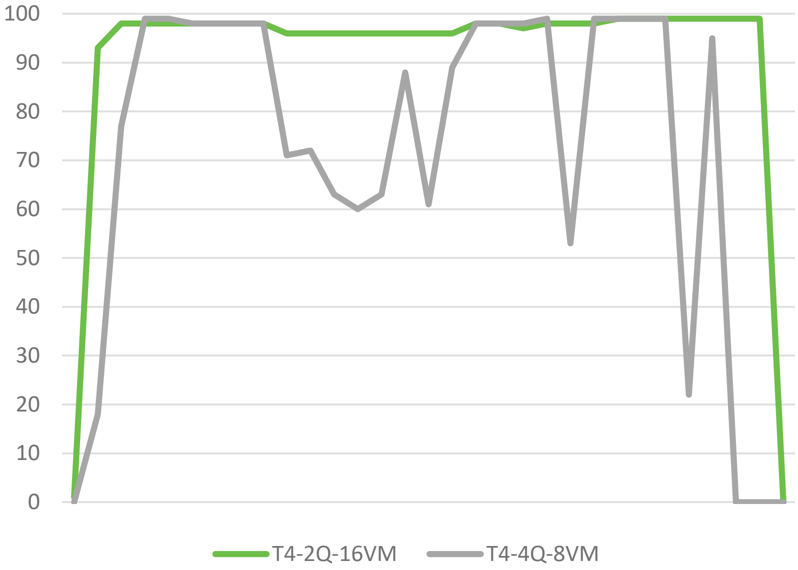
Tying the Benchmarks to CPUs and GPUs

Intel Scalable Family 6140 and NVIDIA Tesla T4*

Intel Xeon 6140 Utilization



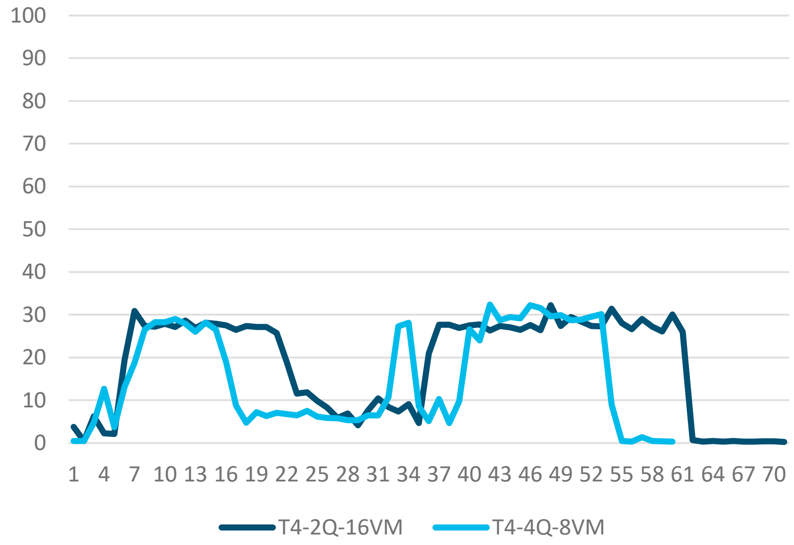
NVIDIA Tesla T4 Utilization



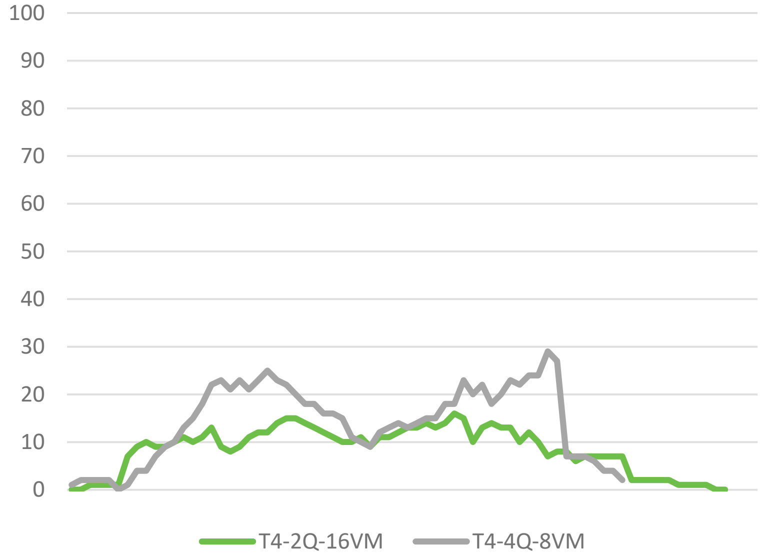
* SPECviewperf 13 Catia Test – ESXi Host Data

Windows 10 (1607) VM with Tesla T4 vGPU*

Perfmon VM Processor Utilization



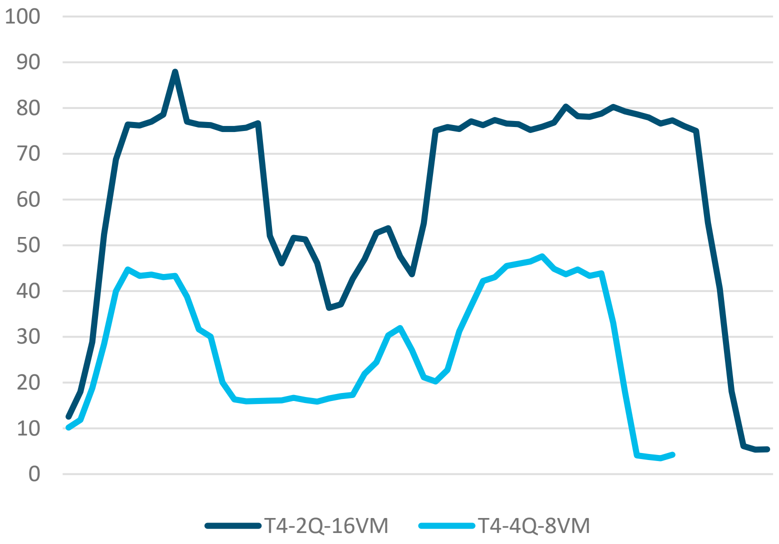
Perfmon VM GPU Utilization



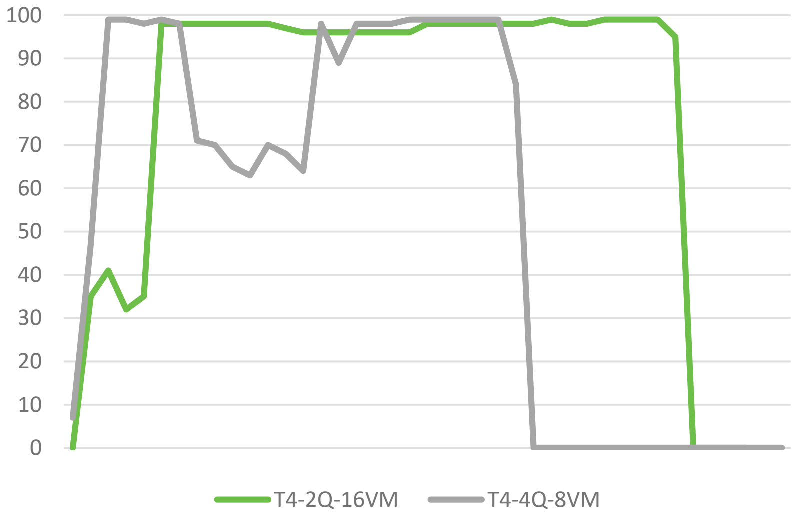
* SPECviewperf 13 Catia Test – Single VM in multiple VM test Perfmon Data with Xeon 6140 host processor

Intel Scalable Family 6136 and NVIDIA Tesla T4*

Intel Xeon 6136 Utilization



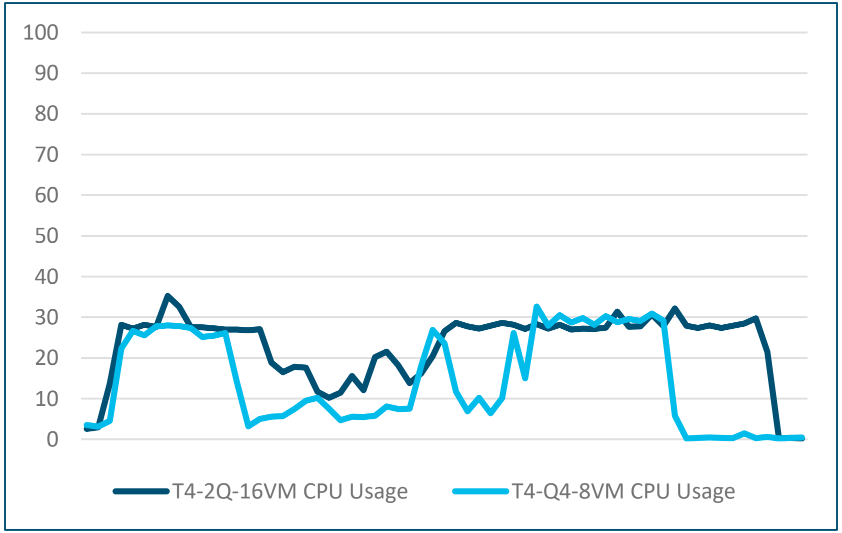
NVIDIA Tesla T4 Utilization



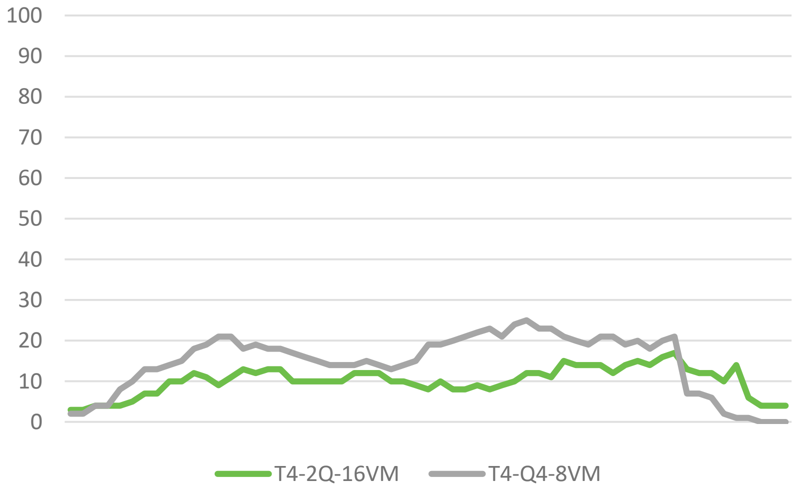
* SPECviewperf 13 Catia Test – ESXi Host Data

Windows 10 (1607) VM with Tesla T4 vGPU*

Perfmon VM Processor Utilization



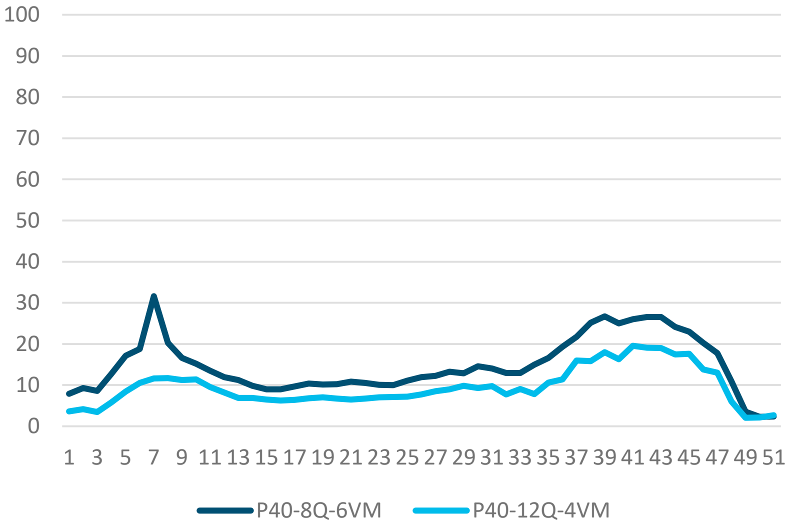
Perfmon VM GPU Utilization



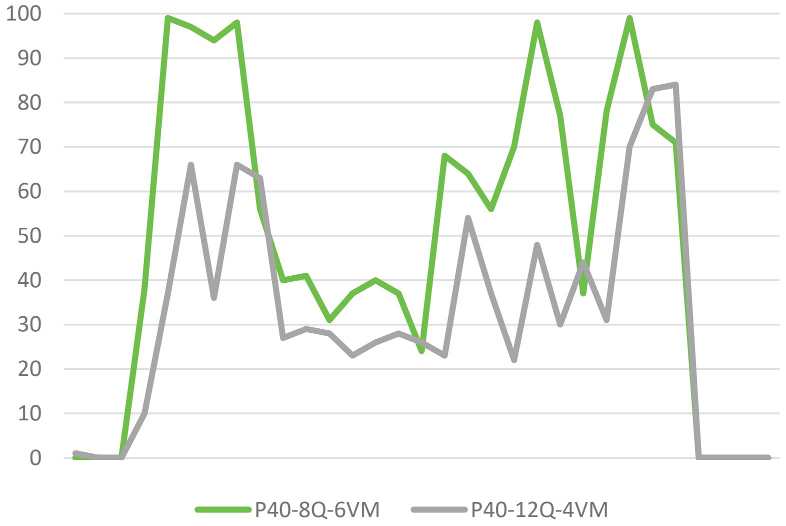
* SPECviewperf 13 Catia Test – Single VM in multiple VM test Perfmon Data with Xeon 6136 host processor

Intel Scalable Family 6140 and NVIDIA Tesla P40*

Intel Xeon 6140 Utilization



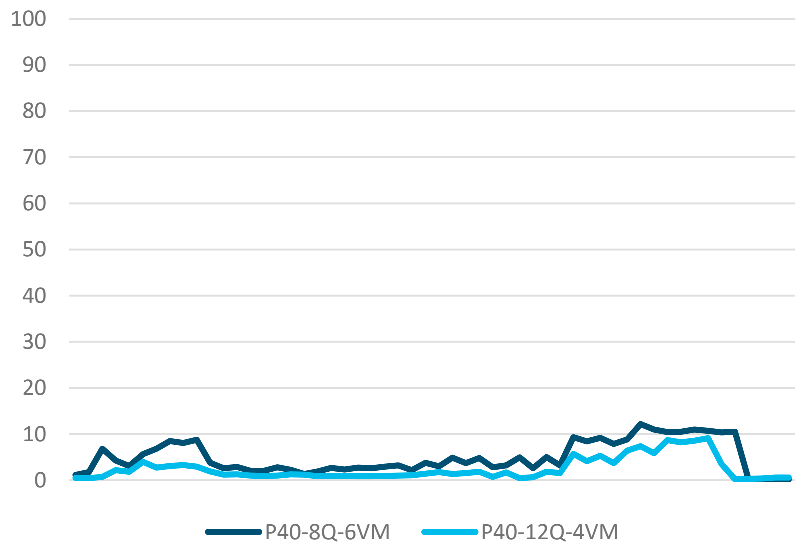
NVIDIA Tesla P40 Utilization



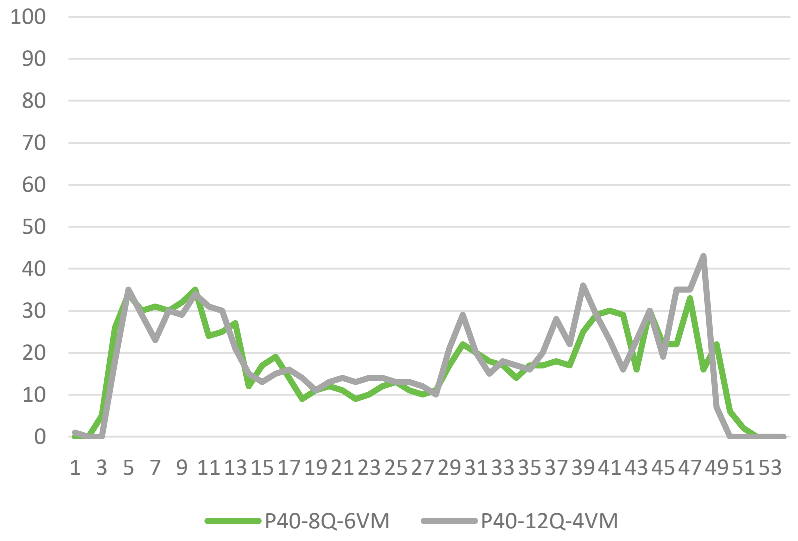
* SPECviewperf 13 Catia Test – ESXi Host Data

Windows 10 (1607) VM with Tesla P40 vGPU*

Perfmon VM Processor Utilization



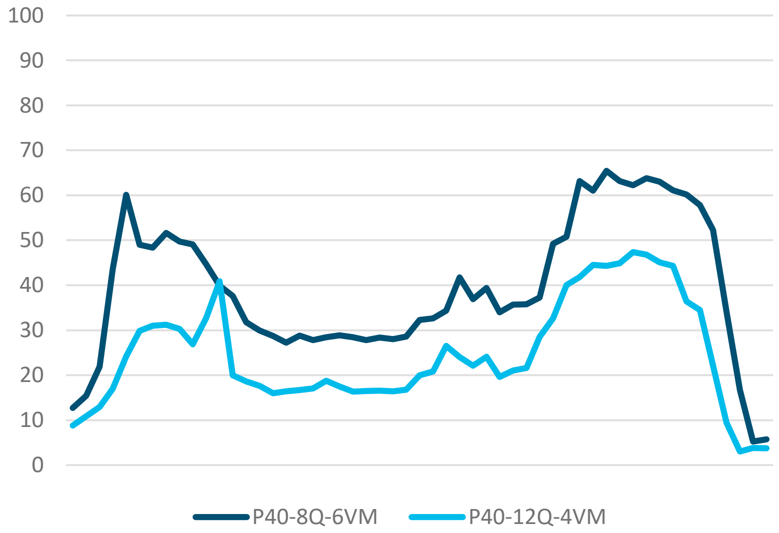
Perfmon VM GPU Utilization



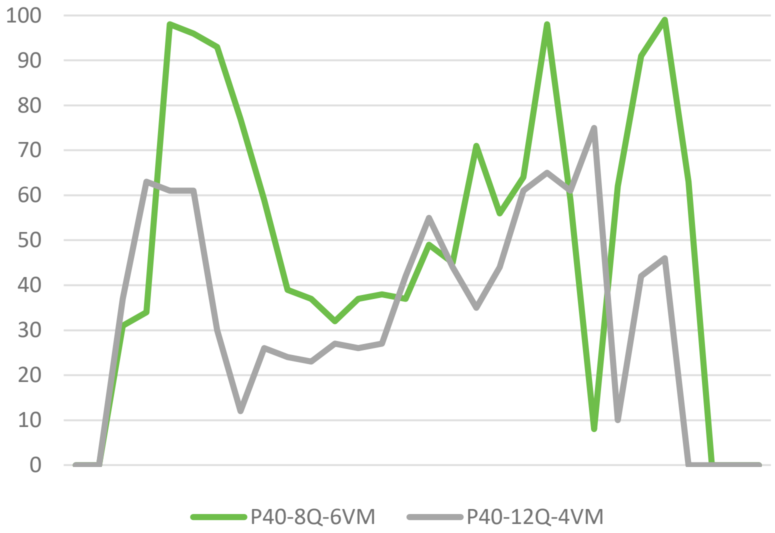
* SPECviewperf 13 Catia Test – Single VM in multiple VM test Perfmon Data with Xeon 6140 host processor

Intel Scalable Family 6128 and NVIDIA Tesla P40*

Intel Xeon 6128 Utilization



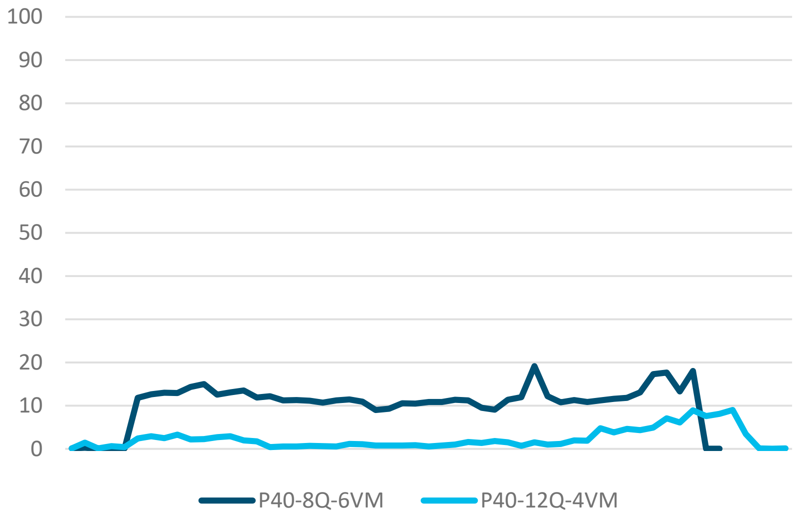
NVIDIA Tesla P40 Utilization



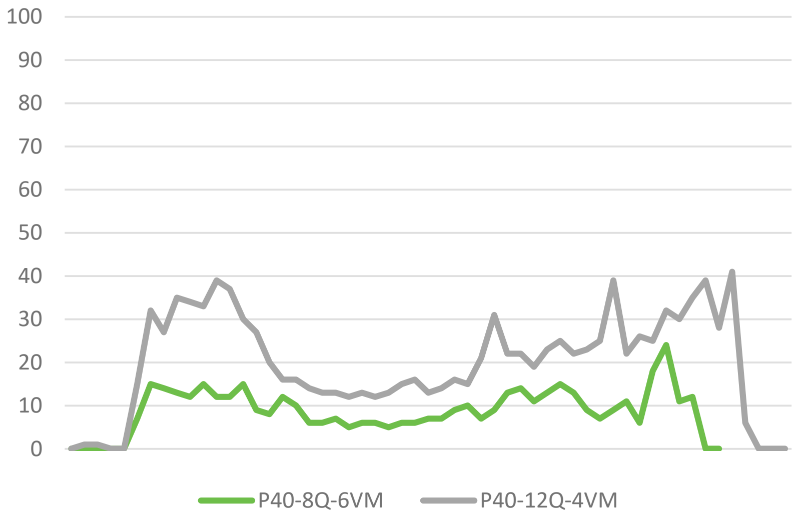
* SPECviewperf 13 Catia Test – ESXi Host Data

Windows 10 (1607) VM with Tesla P40 vGPU*

Perfmon VM Processor Utilization



Perfmon VM GPU Utilization



* SPECviewperf 13 Catia Test – Single VM in multiple VM test Perfmon Data with Xeon 6128 host processor

Sizing for Dassault Apps

Dassault Systemes 3DEXPERIENCE UCS C240 M5 Rack Server starting points*

User type	Equivalent performance	Users/ server	vCPU/ user	Memory/ user	Server CPU	Server memory	NVIDIA GPU	Quadro profile	Storage type	Network
Light	Quadro P1000	32	4	12-16	Intel Xeon 6136	768	Tesla T4 (4)	T4-2Q	Flash	10Gb+
Medium	Quadro P2000	16	4-6	16-32	Intel Xeon 6134	768	Tesla T4 (4)	T4-4Q	Flash	10Gb+
Heavy	Quadro P5000	4-6	8-12	96+	Intel Xeon 6128	768	Tesla P40 (2)	P40-8Q P40-12Q	Flash	10Gb+

*The recommendations above reflect starting points. Customers should perform PoCs to determine optimal configurations for their specific environments. Cisco can help.

Dassault Systemes 3DEXPERIENCE UCS B200 M5 Blade Server Rack Dense starting points*

User type	Equivalent performance	Users/ server	vCPU/ user	Memory/ user	Server CPU	Server memory	NVIDIA GPU	Quadro profile	Storage type	Network
Light	Quadro P1000	12	4	12-16	Intel Xeon 6128	192	Tesla P6 (2)	P6-2Q	Flash	10Gb+
Medium	Quadro P2000	6	4-6	16-32	Intel Xeon 6128	192	Tesla P6 (2)	P6-4Q	Flash	10Gb+
Heavy	Quadro P5000	2-4	8-12	96+	Intel Xeon 6128	192	Tesla P6 (2)	P6-8Q P6-16Q	Flash	10Gb+

*The recommendations above reflect starting points. Customers should perform PoCs to determine optimal configurations for their specific environments. Cisco can help.

The Cisco Lineup

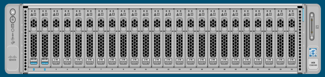
Cisco graphics accelerated Data Center with NVIDIA

Racks

C220 M5

2x NVIDIA P4, T4

C240 M5

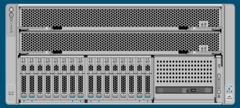


2x NVIDIA V100,P100,P40

2x M10,M60,

6x P4, 4x T4

C480 M5



6x NVIDIA V100,P100, P40, M60

3x M10

C480 ML M5

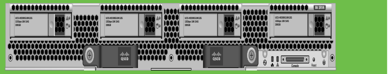
8x NVIDIA V100 32GB NvLINK Interconnect

Blades

B200 M5

2x NVIDIA P6 GPU/blade
Up to 16x per chassis

B480 M5



4x P6 GPU/blade,
Up to 16x per chassis

Hyperconverged

HyperFlex 240C M5

2x NVIDIA V100, P40

2x M10, M60

6x P4

Key takeaways

Keep these things in mind

Understanding the different types of users

There are three key GPU settings:

- GPU scheduler
- NVIDIA profile selection
- Frame rate control
  - NVIDIA Tesla card
  - Desktop Broker

CPU selection is critical

- CPU and GPU work synergistically
- High frequency
- CPU core count

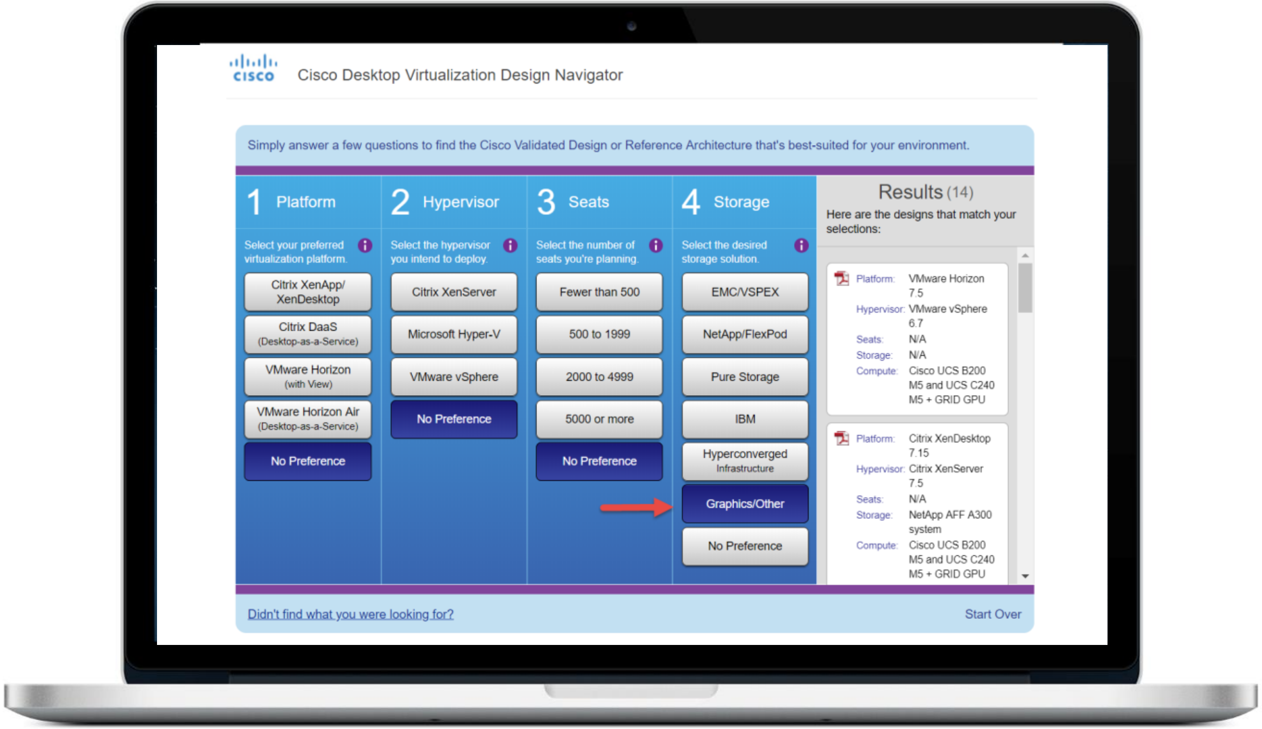
High frequency memory

Resources

VCC Design Navigator

Your source for VDI content

http://cisco.com/go/vdi-cvd



Cisco Desktop Virtualization Design Navigator

Simply answer a few questions to find the Cisco Validated Design or Reference Architecture that's best-suited for your environment.

1 Platform

Select your preferred virtualization platform.

2 Hypervisor

Select the hypervisor you intend to deploy.

3 Seats

Select the number of seats you're planning.

4 Storage

Select the desired storage solution.

Results (14)

Here are the designs that match your selections:

Platform: VMware Horizon 7.5
Hypervisor: VMware vSphere 6.7
Seats: N/A
Storage: N/A
Compute: Cisco UCS B200 M5 and UCS C240 M5 + GRID GPU

Platform: Citrix XenDesktop 7.15
Hypervisor: Citrix XenServer 7.5
Seats: N/A
Storage: NetApp AFF A300 system
Compute: Cisco UCS B200 M5 and UCS C240 M5 + GRID GPU

Didn't find what you were looking for?

Start Over

Q & A

cisco