VDI Evolution at the speed of GRID – VDI 2.0 IS here!

Shawn Kaiser – TSA Cisco Systems

Jason Marchesano – TSA Cisco Systems

About us

Shawn and Jason are both Technical Solutions Architects with Cisco Systems focused on Server and Desktop Virtualization.

They have been in the Virtualization industry since ESX was first introduced. Both have been end customers of the technology, consultants for other customers with Virtualization initiatives and now work with directly with Cisco customers and the Cisco UCS Product Development on next generation architectures.

The case for GPU with Desktop Virtualization

- Why is GPU needed?
- Technology for GPU in Desktop Virtualization
- Methodology for implementing GPU
  - Virtual Workstations
  - General Purpose GPU (gpGPU)
- VDI 2.0, better with Cisco UCS

Why GPU in Desktop Virtualization?

VDI 2.0 is Here!

Evolving to VDI 2.0

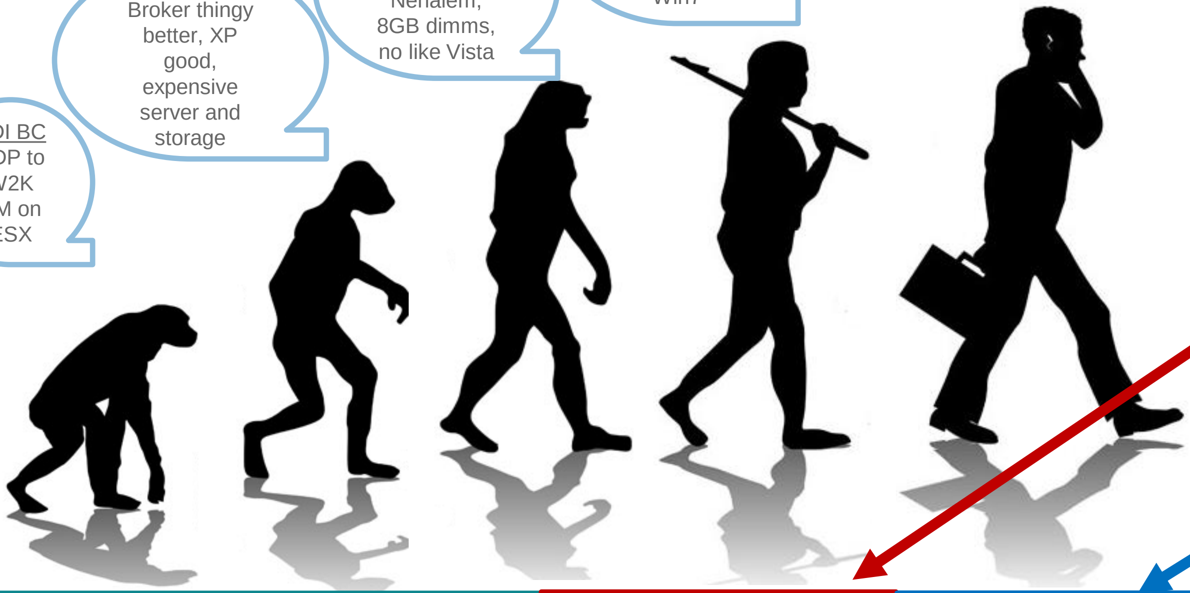
VDI BC
RDP to W2K VM on ESX

VDI .5
Broker thingy better, XP good, expensive server and storage

VDI 1.0
Thin prov, non-pers, Nehalem, 8GB dimms, no like Vista

VDI 1.1
AFA good, cheaper Servers, users want Win7

VDI 2.0
Multiple monitors, multimedia, 3D accelerated devices: End users are spoiled with desktops while I stumbled through VDI 1.0 and 1.1. GPU for VDI is here to save the day!



Problem: Most VDI Infrastructure is Here

Users and Apps are Here

2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

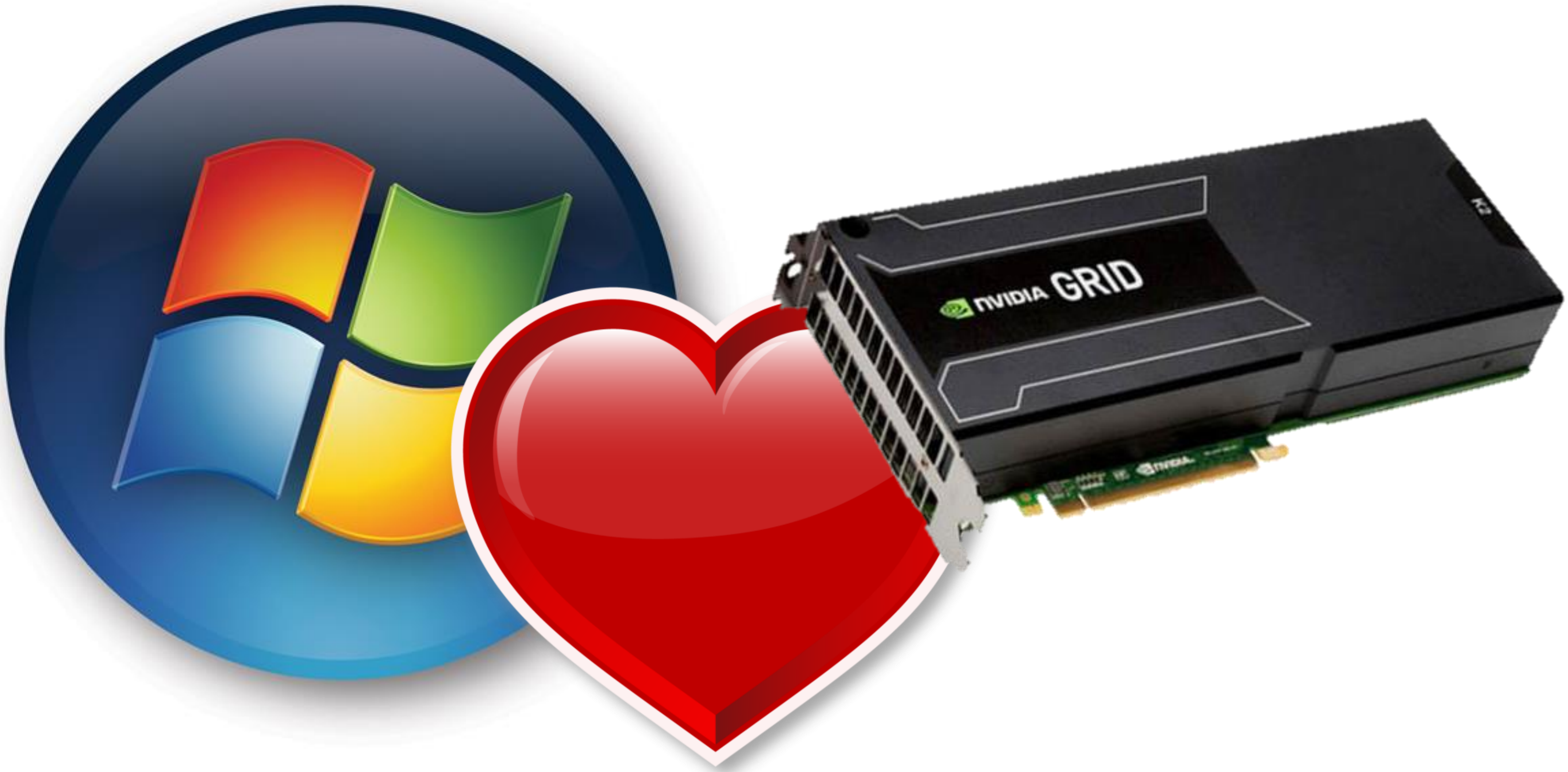
Want More info?

Google: "VDI 2.0 is knocking"

What users want

- VDI 1.0
  - Storage Architecture immaturity
  - Cost woes
  - *JUST GET IT TO WORK! mentality*

- VDI 2.0
  - Users dictate GO/NO GO
  - Needs to look/feel like desktop

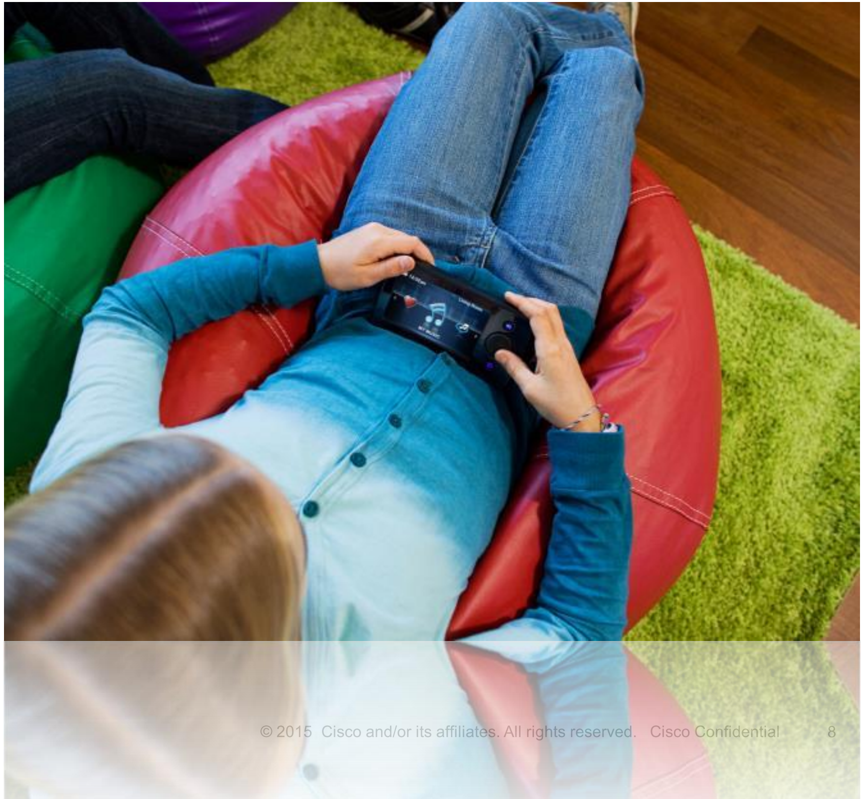

NVIDIA GRID

Isn't a slow cell phone frustrating?

Users Expect Fluid Performance

Cell Phones and Tablets, Laptops and PCs, even automobile displays:

End Users are surrounded by many devices that are GPU driven

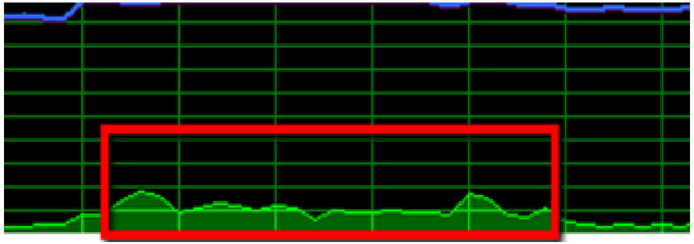


Windows Hearts GPU

- Windows Vista introduced Aero Desktop Experience…GPU required
- Windows 10 will likely be next corporate standard, Doubles idle GPU ram requirements
- ISV's have been writing to the DirectX / OpenGL API for years
  - Office 2013, Internet Explorer, Chrome, Firefox All GPU accelerated

Science Experiment

By default, hardware graphics acceleration is enabled in Office 2013

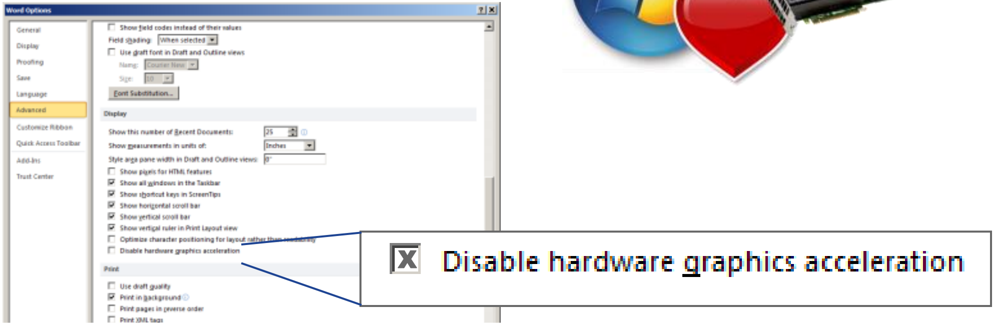


Scrolling in Word with GPU Acceleration turned ON =
10-15% CPU Utilization

…GPU Co-processing reduces CPU Burden for screen updates!

Science Experiment

But eliminate hardware graphics acceleration….



☐ Show field codes instead of their values

Field shading: When selected

☐ Use draft font in Draft and Outline views

Name: Courier New

Size: 10

Font Substitution...

Display

Show this number of Recent Documents: 25

Show measurements in units of: Inches

Style area pane width in Draft and Outline views: 0"

☐ Show pixels for HTML features

☑ Show all windows in the Taskbar

☑ Show shortcut keys in ScreenTips

☑ Show horizontal scroll bar

☑ Show vertical scroll bar

☑ Show vertical ruler in Print Layout view

☐ Optimize character positioning for layout rather than readability

☐ Disable hardware graphics acceleration

Print

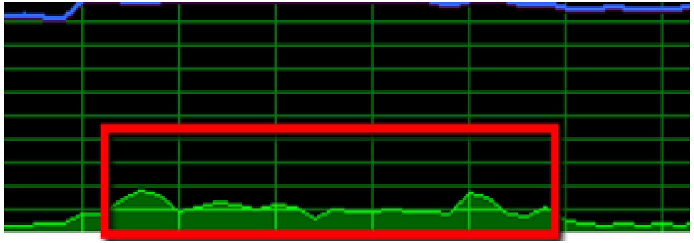
☐ Use draft quality

☑ Print in background

☐ Print pages in reverse order

☐ Print XML tags

☒ Disable hardware graphics acceleration



Scrolling in Word with GPU Acceleration turned ON =
10-15% CPU Utilization

…and CPU burden is increased dramatically

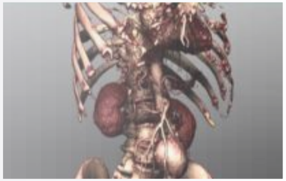
GPU Requirement for VDI User Profile

DESIGNER
Graphics and Media Professionals, Design Engin

CATIA, CS6, Inventor

POWER USER
Financial Analysts, Trad
Design Reviewers

works, Adobe
eaver, Medical
ing Showcase



KNOWLEDGE WORK
Office workers, productivity and
line-of-business workers

MS Office, Photoshop

Regardless of “profile”…

All users can benefit from a GPU enhanced experience

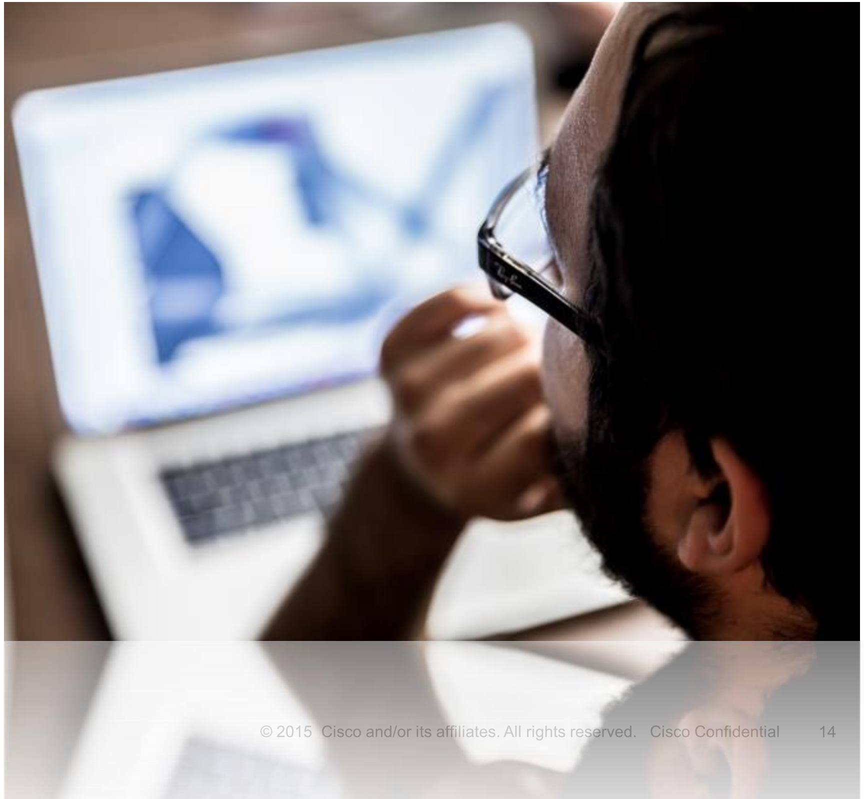
Simplifying the world of 3D Accelerated Virtual Desktops

3D Accelerated Virtual Desktop Tech

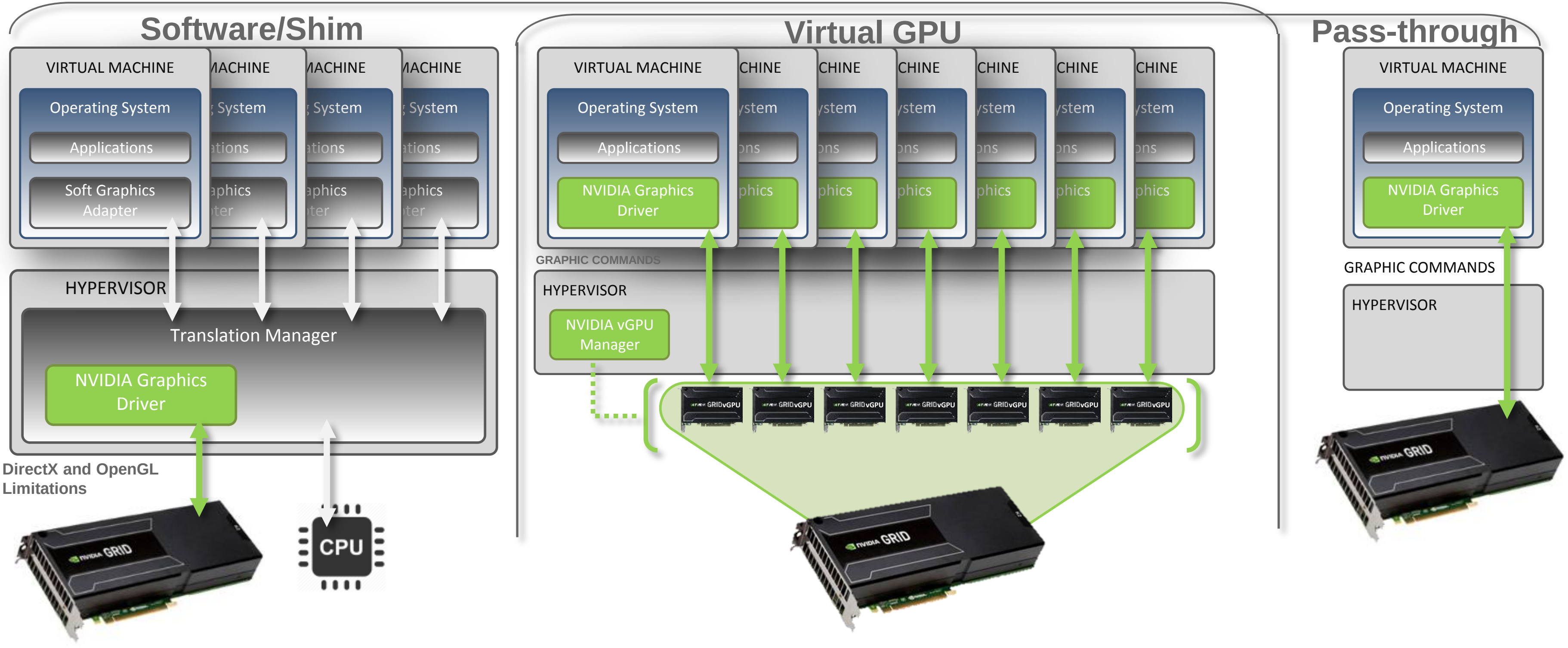
is *Evolving* for the better.

Multiple options for various use cases provides various levels of performance and consolidation.

Hypervisors and brokers are embracing NVIDIA GRID technology.



Its about getting the GPU to the VM…



GPU support for VDI Profile

Vendor	GPU Pass-Through	GPU Soft-Sharing	NVIDIA GRID Virtual GPU
CiTRIX	✓	XenApp 6.5 on Windows Server	✓
vmware	✓ (vDGA)	✓ (vSGA)	✓ (vSphere 6)
Microsoft	✓	✓ (RemoteFX)	✗

Hypervisors, Brokers and Apps

- How to implement a 3D accelerated desktop will vary based on hypervisor, broker, and application needs
- “Layer 8” issues as well
  - What will the customer allow in their datacenter?

Making Sense of the Madness…

Implementing GPU in Desktop Virtualization

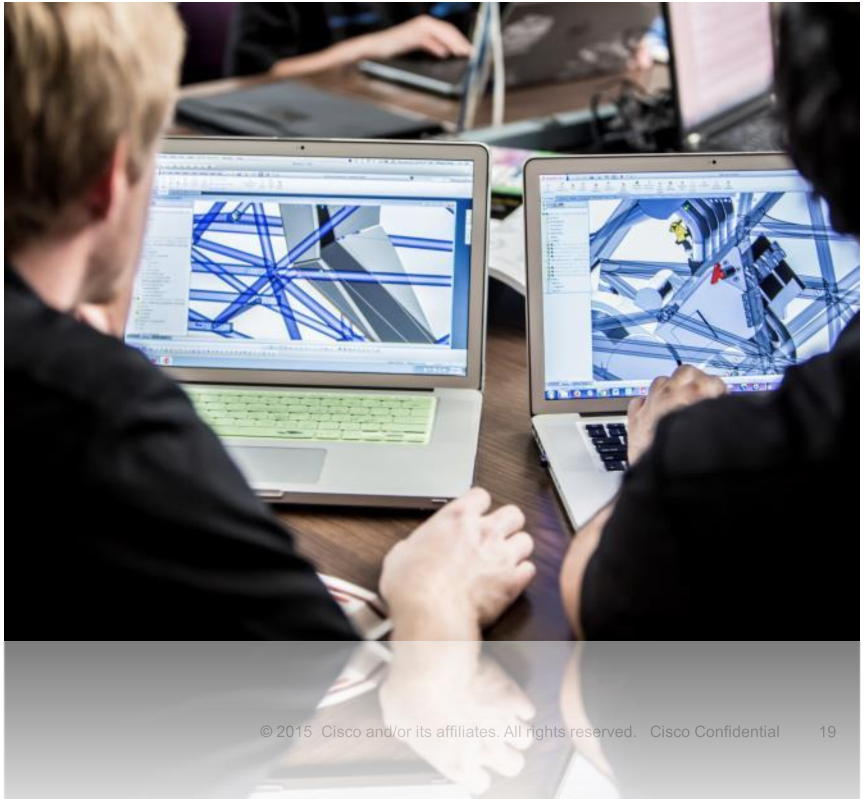
Sizing Philosophy 101

Know your User…

Know your App…

Know what equipment you are replacing…

…architect accordingly!



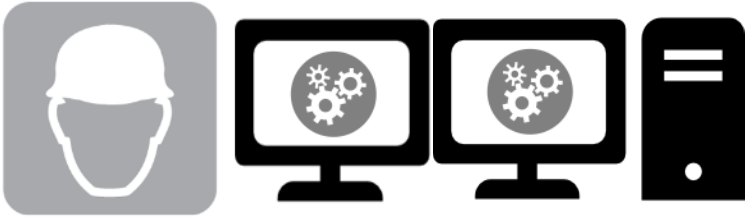
Simplify your deployment by determining category of your end use case…

- Task / Power User (General Purpose)
  - Office 2013
  - Browser Based Applications
  - Client / Server database applications

- Power User / Designer (Virtual Workstation)
  - Catia
  - SOLIDWORKS
  - AutoCAD

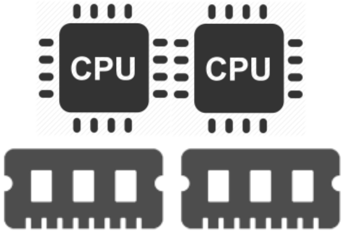
What equipment is being replaced?

- General Purpose Desktop
  - Low cost dual core CPUs
  - Spinning Disk
  - Low cost GPU/APU integrated into CPU
  - Single or dual lower resolution displays
- Professional Workstation
  - Multi Socket / Multi core CPUs
  - FLASH Storage
  - Dedicated Quadro GPUs
  - Multiple high resolution displays

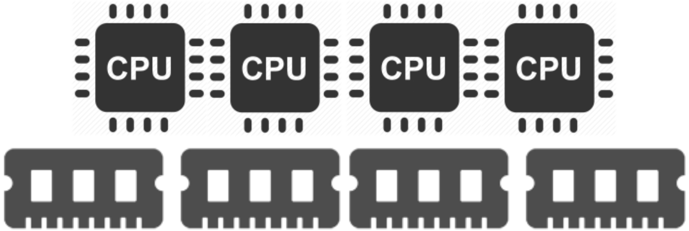


Consider Virtual Desktop VM Sizing

- General Purpose Virtual Desktop
  - 1-2 vCPUs
  - 2-8 GB of RAM



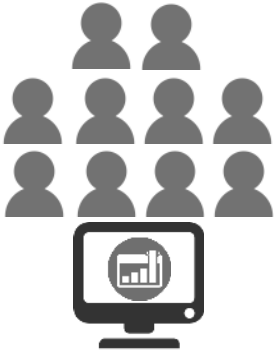
- Virtual Workstation
  - 2-6 vCPUs
  - 8-32 GB of RAM



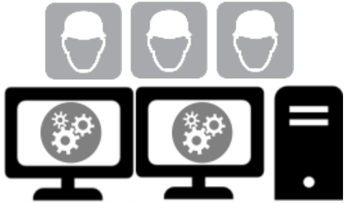
Determine Server Oversubscription

Manage delivered performance by limiting resource contention!

- General Purpose Virtual Desktop
  - Up to 10:1 vCPU:pCPU overcommit
  - 64 - 128 Users per server depending on GPU sharing technology



- Heavy Duty Virtual Workstation
  - Between 1:1 and 2:1 vCPU:pCPU overcommit
  - Depending on application, 4-16 users max per server



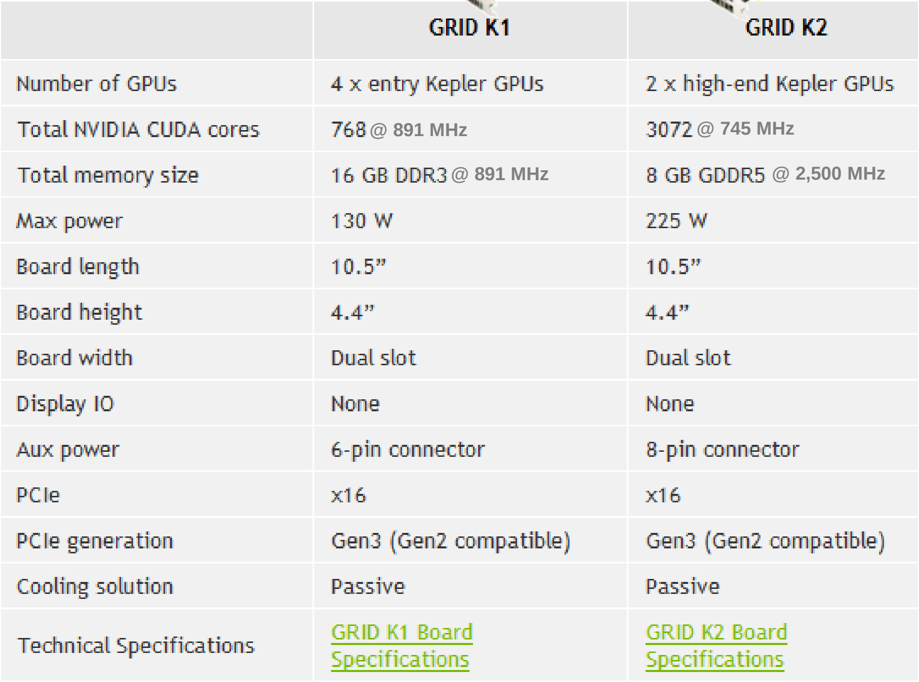
Which NVIDIA GRID Card?

General Purpose Virtual Desktop

- NVIDIA GRID K1
- vSGA – up to 128 users
- vGPU – pick the right vGPU profile

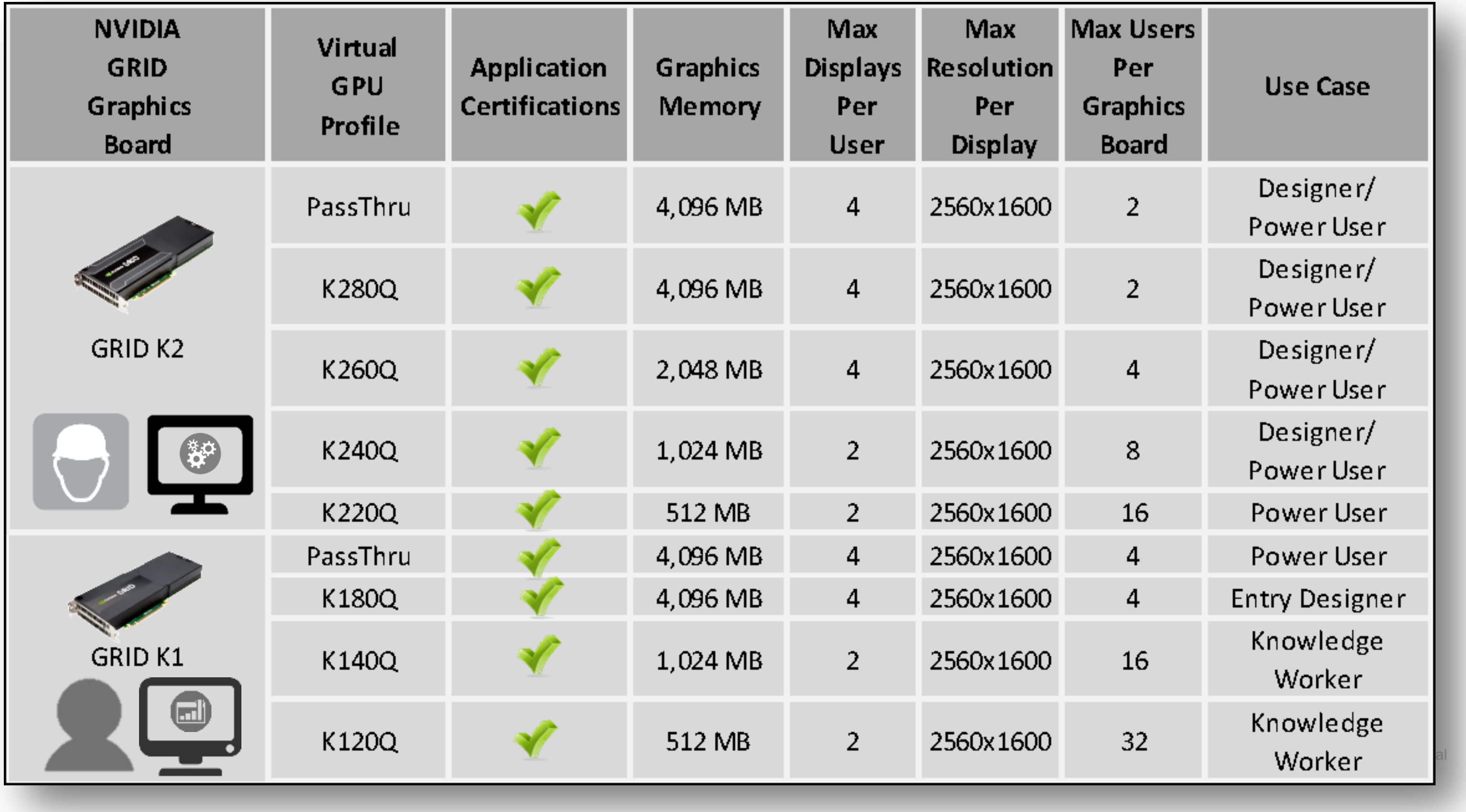
Virtual Workstation

- NVIDIA GRID K1 or K2
- vGPU – pick the right vGPU profile
- vDGA – full GPU power/ CUDA support



	GRID K1	GRID K2
Number of GPUs	4 x entry Kepler GPUs	2 x high-end Kepler GPUs
Total NVIDIA CUDA cores	768 @ 891 MHz	3072 @ 745 MHz
Total memory size	16 GB DDR3 @ 891 MHz	8 GB GDDR5 @ 2,500 MHz
Max power	130 W	225 W
Board length	10.5”	10.5”
Board height	4.4”	4.4”
Board width	Dual slot	Dual slot
Display IO	None	None
Aux power	6-pin connector	8-pin connector
PCIe	x16	x16
PCIe generation	Gen3 (Gen2 compatible)	Gen3 (Gen2 compatible)
Cooling solution	Passive	Passive
Technical Specifications	GRID K1 Board Specifications	GRID K2 Board Specifications

NVIDIA vGPU Profiles



NVIDIA GRID Graphics Board	Virtual GPU Profile	Application Certifications	Graphics Memory	Max Displays Per User	Max Resolution Per Display	Max Users Per Graphics Board	Use Case
GRID K2	PassThru	✓	4,096 MB	4	2560x1600	2	Designer/ Power User
	K280Q	✓	4,096 MB	4	2560x1600	2	Designer/ Power User
	K260Q	✓	2,048 MB	4	2560x1600	4	Designer/ Power User
	K240Q	✓	1,024 MB	2	2560x1600	8	Designer/ Power User
	K220Q	✓	512 MB	2	2560x1600	16	Power User
GRID K1	PassThru	✓	4,096 MB	4	2560x1600	4	Power User
	K180Q	✓	4,096 MB	4	2560x1600	4	Entry Designer
	K140Q	✓	1,024 MB	2	2560x1600	16	Knowledge Worker
	K120Q	✓	512 MB	2	2560x1600	32	Knowledge Worker

Do not forget the Network!

10 Gb/s fabric is now the norm

Watch for bottlenecks between Desktop Virtualization farms and end user data. A Virtual Workstation with very large application datasets can perform better than physical when end to end 10Gb fabric is available

Remote screen refresh will be limited by WAN connectivity

Desktop Virtualization 2.0: Simplified with Cisco Unified Computing

A great virtual desktop experience is not just about the physical hardware behind it..

Infrastructure should simplify your day to day management overhead…

Cisco Unified Computing Solution greatly simplifies the operations of a Desktop Virtualization Solution.

Cisco Unified Computing System

A differentiated/revolutionary approach

Simplified Architecture

- Networking with fewer components
- Lower cost and easier scaling
- Fewer management touch points
- Stateless: any resource, any time
- Better TCO/ROI

Unified Management

- Faster deploy/ provision
- Unification leads to reduced complexity
- Centralized Firmware Provisioning
- Management via a single interface

Higher Performance

- Brings out the best of x86 architecture
- Optimized resource utilization for compute, networking, and management
- Low latency network fabric

Scale

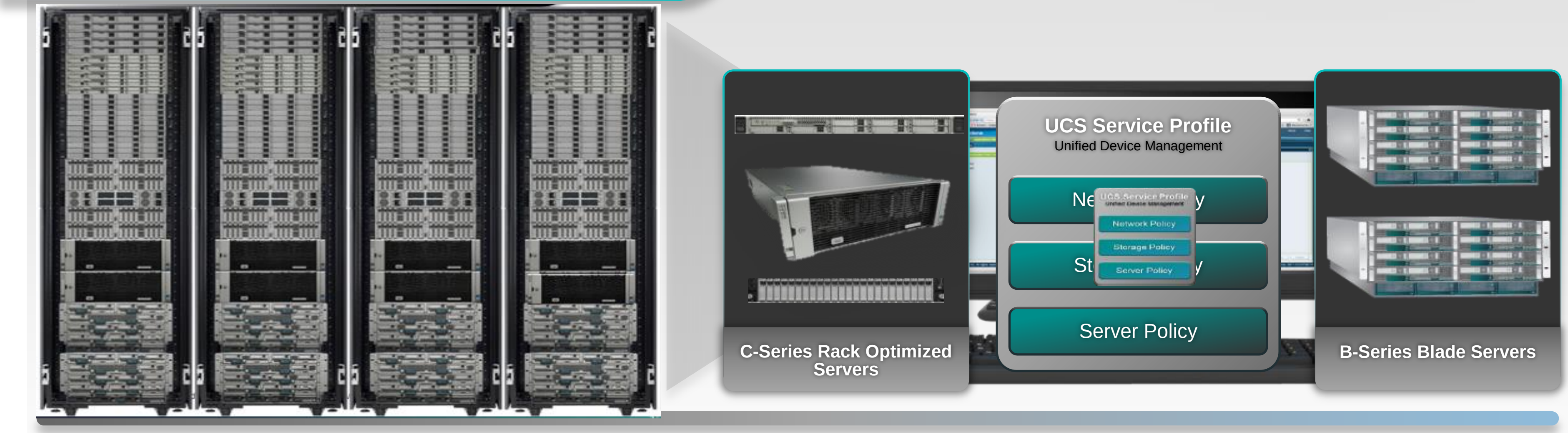
- Ultimate Scalability
- Enhanced design capability
- Designed for the future, today

Unified Management

Extending Benefits of UCS Manager to Rack Servers

Only UCS Manages Blades and Racks as One

The **Only** Vendor With A Single Unified Management For Blade and Rack Servers



- A major market transformation in unified server management
- Benefits of UCS Manager and Service Profiles brought to rack optimized servers

- New Nexus Fabric Extender topology reduces cost, increases scale of rack server connectivity within Unified Computing
- Add capacity without complexity

UCS Unified Management Blade & Rack

Cisco is the **Only** vendor in the industry to architect a common management engine using model and policy-based control to fully configure and create templates for both *Blade* and *Rack* Servers.

Cisco UCS C240 M4

Ideal Platform for GPU accelerated Desktop Virtualization

- Supports 2 Grid K1 or K2 or combination there of…
- Up to 36 Intel E5v3 CPU Cores
- Up to 756 GB of RAM
- 1, 10, or 40 Gb/s connectivity

Certified to support 2 Grid cards with Intel’s top bin, 145W CPUs

GPU Density highest with Rack Mounts

More GPUs with less infrastructure per Rack U than blades!

Still completely managed by UCS Manager!

What if I need more GPU?

Introduction to Magma EB3600

PCIe Gen3 Expansion Chassis

- Up to 9 double width GPU’s
- Share between multiple systems
- Increases per Server GPU Density

Qualification in process on C240 M4

PCIe Expansion balances GPU and Compute

More GPUs with less infrastructure per Rack U than blades!

More cost effective GPU when density is key.

Virtualization Optimized with VM-FEX

Performance

Up to 50% increase
in Application performance

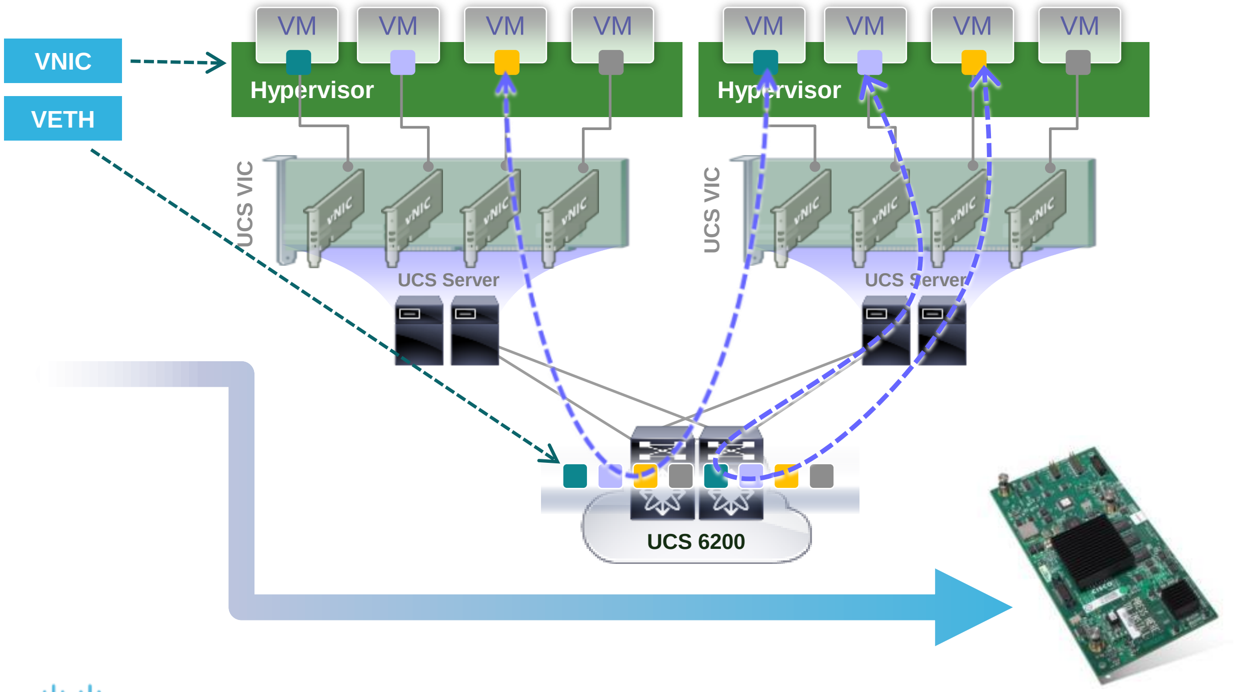
Low
Latency

Up to 67% reduction
in Application latency

Deterministic
Delivery

Near linear deterministic Application delivery with scale

Desktop Virtualization Performance Optimized with VM-FEX



VM-FEX Basics

- Fabric extender for VMs
- Hypervisor vSwitch removed
- Each VM assigned a PCIe device
- Each VM gets a virtual port on physical switch

VM-FEX: One Network

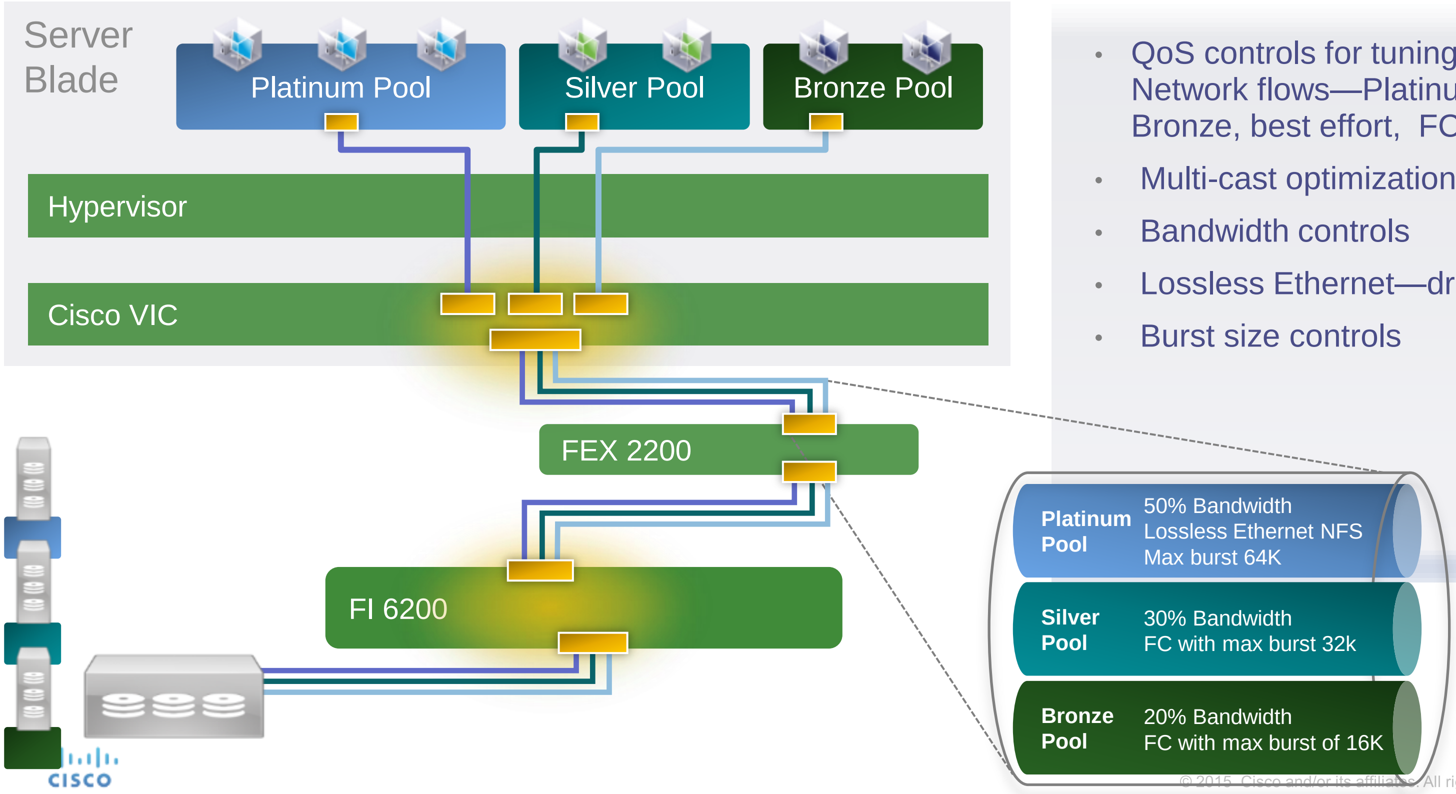
- Collapses virtual and physical switching layers
- Dramatically reduces network management points by eliminating per host vSwitch
- Virtual and physical traffic treated the same

Host CPU Cycles Relief

- Host CPU cycles relieved from VM switching
- I/O throughput improvements

Cisco UCS with VM-FEX:
Virtual Desktop Prioritization and QoS Pools

- QoS controls for tuning Storage and Network flows—Platinum, Gold, Silver, Bronze, best effort, FC QoS Classes
- Multi-cast optimizations
- Bandwidth controls
- Lossless Ethernet—drop/no drop
- Burst size controls

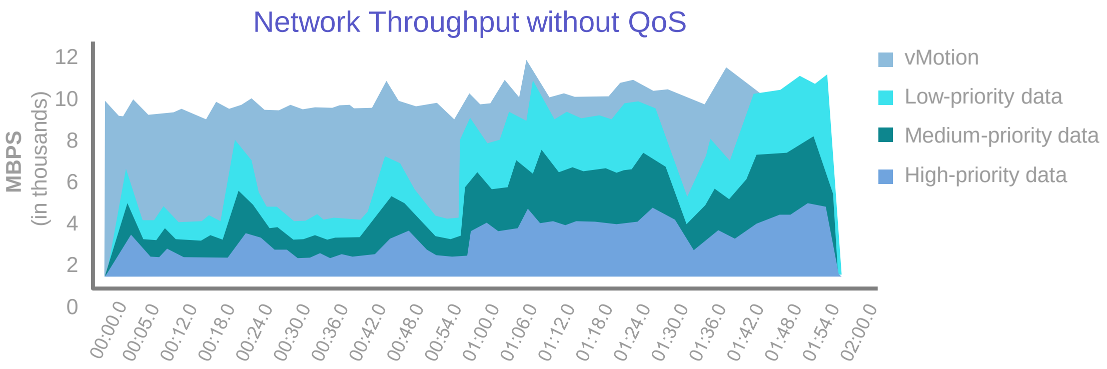


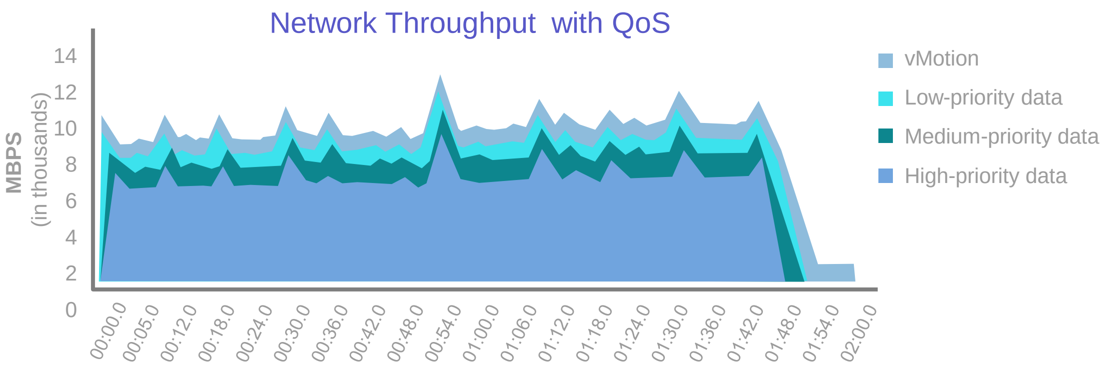
Platinum Pool 50% Bandwidth Lossless Ethernet NFS Max burst 64K

Silver Pool 30% Bandwidth FC with max burst 32k

Bronze Pool 20% Bandwidth FC with max burst of 16K

VDI Flow Prioritization and QoS Pools





- User experience and SLA association to the virtual desktop
- Prioritization among multiple virtual desktop pools
- Consistent virtual desktop behavior with vMotion, backup and other data center actions
- Burst controls, and other traffic shaping controls
- Separation of cluster management traffic from desktop traffic
- Up to 80 Gb/s bandwidth per server to prevent HOL blocking

High Performance Desktop Virtualization 2.0 Is Here!

Remember These Points…

Current generation Desktop OS's require GPU support.

End users demand a modern, full desktop experience.

Understand and apply appropriate NVIDIA GRID technology.

Cisco UCS Leads with a fully managed, performance optimized solution.