White Paper

Graphics Accelerated Virtual Desktops and Applications for High-Technology Computing

FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID

End-User Computing Technical Marketing, NetApp

February 2015 | WP-7212

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

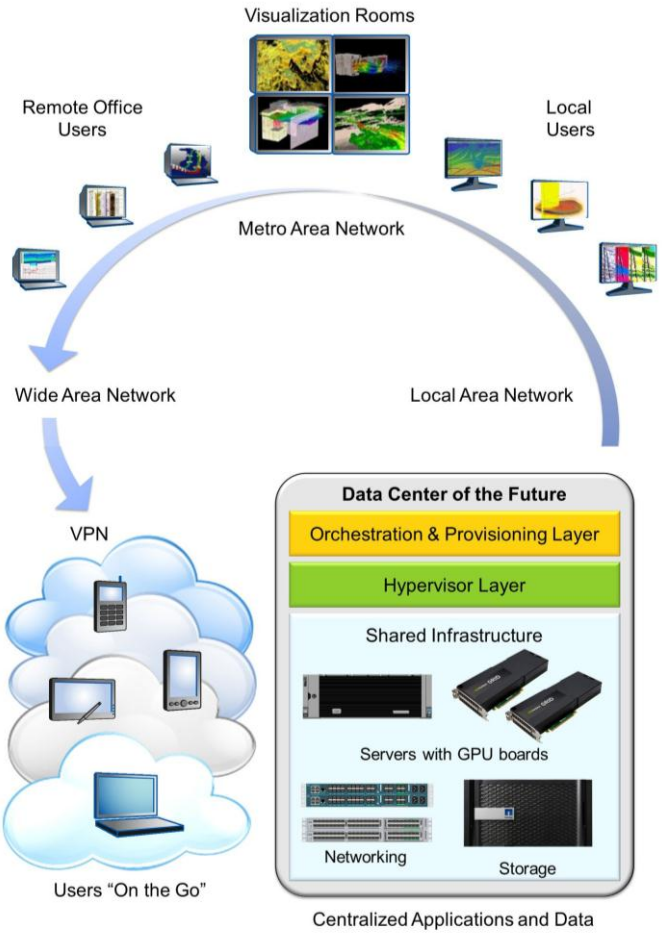
1 A New Approach to Visualization: Mobile, Secure, Cost Effective

Allowing a mobile workforce to visualize graphics-intensive data wherever they happen to be is becoming essential for many industries, such as oil and gas, manufacturing, healthcare, and other high-technology use cases. Normally, such visualization requires either local dataset copies, which can be difficult and potentially risky to provide, or remote visualization. Up to now, remote visualization solutions have been complicated to deploy and manage, and they haven't provided the quality or responsiveness that users expect.

NetApp and Cisco partnered with Citrix and NVIDIA to address the demanding visualization requirements across a number of use cases. Combining the proven NetApp® FlexPod® Datacenter architecture with the advanced virtualization and visualization technologies from Citrix and NVIDIA results in a solution that makes visualization accessible to everyone who needs it, including users with thin devices such as tablets. This new solution, FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID, was carefully architected to address the challenges of graphics-intensive visualization workloads.

As shown in Figure 1, when FlexPod Datacenter is deployed with Citrix XenDesktop and NVIDIA GRID, a single, reliable infrastructure is able to address a full spectrum of visualization needs, including those of local users, visualization rooms, remote office users, and workers in the field.

Figure 1) FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID.



2 High-Performance Visualization Challenges

High-performance visualization requirements are significantly different from those of typical IT infrastructures. It's not uncommon to run hundreds of different applications with a broad mix of I/O needs that can include both sequential and random data access, potentially higher I/O requirements compared to traditional desktop virtualization deployments, extremely large file sizes, and huge numbers of smaller files.

Workflows continue to evolve because of the rapid growth in dataset size and the need to integrate diverse, multidisciplinary data types to more accurately understand and interpret the data. Feeding all this data to visualization workstations outside the data center has become increasingly difficult. Network connections lack the necessary bandwidth for real-time operations, and workflows increasingly span organizational boundaries. Continued reliance on visualization workstations creates significant challenges, including:

- Copying data to a workstation to facilitate visualization, which is sometimes the only option and can take several hours, wastes valuable time during transfers.
- Data on local workstations may not be protected by backup or disaster recovery, risking data loss and/or security breaches.
- Storing data on workstations outside the data center increases security risks.
- Moving data outside the data center or across international boundaries may violate regulations and corporate governance requirements.
- Workstation failures can have a big impact on the productivity of engineers, designers, scientists, doctors, and other visualization users.
- Software licensing costs per seat can be extremely high, while productivity per license may be low.
- Collaboration, both local and remote, often requires additional data movement, creating delays of hours or even days before collaboration can begin.
- Bringing a new workstation online to support a new user can take weeks.
- Setting up workstations for workers in remote locations and supplying them with data create additional challenges and additional expense.

The upshot is that significant opportunities exist to increase productivity, enhance collaboration, decrease risks, simplify the overall IT environment, and decrease costs.

3 FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID

The FlexPod Datacenter with Citrix and NVIDIA solution brings together best-in-class components to overcome the challenges just discussed. Data is centralized in the data center, which eliminates the need for data copies, increases security by reducing risk, decreases management overhead, and increases productivity. Anyone who needs visualization can request a session and begin working in a matter of moments.

FlexPod Datacenter with Citrix and NVIDIA extends the familiar concept of virtual desktop infrastructure (VDI) to encompass real-time, 3D graphics capabilities. An authorized user simply requests a desktop or application session. Data is processed and rendered within the data center, and the results of visualization are transmitted over the network to the user.

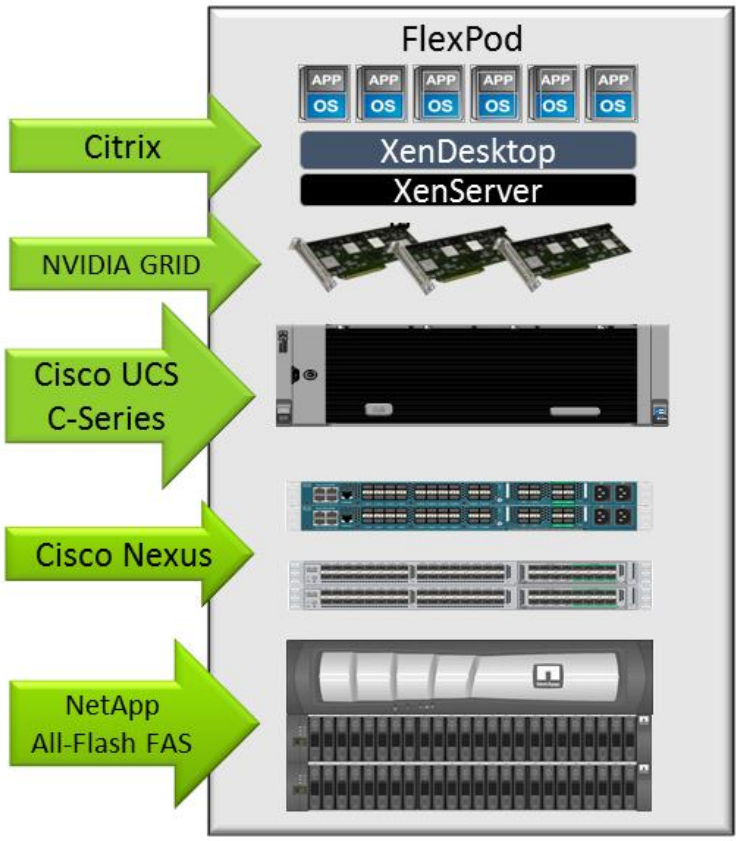
3.1 Proven FlexPod Architecture

FlexPod is a proven data center solution from NetApp and Cisco, offering a flexible, shared infrastructure that easily scales to support growing workload demands without affecting performance. FlexPod has been widely used in desktop virtualization deployments for the last five years. By leveraging the FlexPod architecture, this solution delivers the full benefits of FlexPod, including:

- **Faster deployment with less risk.** The prevalidated design means that you can get FlexPod up and running quickly and deploy applications more quickly.
- **Ability to scale up and out.** Unlike other integrated infrastructure solutions, FlexPod grows and adapts to meet your changing needs.
- **Investment protection.** The scalability and flexibility of FlexPod give the solution a longer life. You can easily repurpose components to address changing needs.
- **Cooperative support.** NetApp, Cisco, and Citrix partnered to provide cooperative support. This approach resolves 98% of support issues on first contact.

In addition to these benefits, each component of the FlexPod Datacenter with Citrix and NVIDIA solution delivers specific benefits for graphics-intensive workflows. The full solution consists of NetApp FAS flash storage, Cisco Unified Computing System™ (Cisco UCS®) C-series servers with installed NVIDIA GRID K2 graphics processing units (GPUs) for compute and visualization, Cisco Nexus® for networking, and Citrix XenDesktop HDX 3D Pro for GPU-aware virtualization.

Figure 2) FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID solution components.



3.2 NetApp FAS Storage

NetApp FAS storage reduces overall storage costs while delivering the necessary I/O performance for desktop virtualization and demanding visualization applications. FAS storage supports both all-flash and hybrid storage configurations to create an optimal storage platform for your needs.

The scale-out architecture of NetApp FAS storage and the NetApp clustered Data ONTAP® operating system can flexibly adapt to a variety of I/O workloads. To deliver the necessary high IOPS and throughput, along with low latency, NetApp typically recommends all-flash configurations. All-flash FAS

nodes can be combined in the same scale-out cluster with hybrid (HDD plus flash) storage nodes so the data can be nondisruptively moved between the all-flash and hybrid configurations. You can also nondisruptively replicate or back up your VDI environment (from SSD storage) to more economical HDD storage on other nodes.

NetApp FAS storage offers features that are extremely useful in desktop virtualization environments, simplifying management, increasing availability, and reducing the total amount of storage needed, which in turn makes flash storage cost efficient:

- **Storage efficiency.** Reduce total capacity requirements with deduplication, compression, and thin provisioning.
- **Space-efficient cloning.** Desktop virtualization environments often have many copies of the same operating system and application data. The NetApp FlexClone® capability allows you to almost instantly create clones to support additional desktops. These clones consume additional storage only as changes are made.
- **Integrated data protection.** Full data protection and disaster recovery features allow you to protect critical data assets as well as nondisruptively test and provide disaster recovery.
- **Nondisruptive operations.** Load balancing, upgrading, and maintenance can be performed without taking data offline.
- **Quality of service (QoS).** Storage QoS allows you to limit the storage resources that a particular workload can consume, helping manage noisy neighbor situations.

For a detailed overview of the capabilities of NetApp all-flash FAS storage for Citrix XenDesktop and XenApp deployments, refer to TR-4342: NetApp All-Flash FAS Solution for Persistent and Nonpersistent Desktops with Citrix.

NetApp is the leading data storage provider for key verticals, such as upstream oil and gas, manufacturing, and healthcare industries, and NetApp has forged close relationships over many years with the leading independent software vendors so that NetApp storage systems perform well with important applications.

3.3 Cisco UCS and Cisco Nexus

Cisco UCS unites computing, networking, storage connectivity, and virtualization in a single cohesive system that meets the unique demands of desktop virtualization and 3D visualization. Cisco UCS integrates computing resources with Cisco Nexus switches and a unified I/O fabric, which identifies and handles different types of network traffic, including storage I/O, streamed desktop traffic, management, and access to the applications. All Cisco UCS servers are stateless. Service profiles automate the build process for each server and simplify failure recovery.

Cisco UCS C240 M4 rack servers are typically chosen for the FlexPod Datacenter with Citrix and NVIDIA solution. These servers feature extended memory for faster rendering, support of bigger datasets, more desktops per server, and low latency. This server supports up to two double-wide GPUs in a slim 2U form factor. GPUs can be used to perform parallel tasks (GPU computing) as well as visualization.

3.4 NVIDIA GRID K2

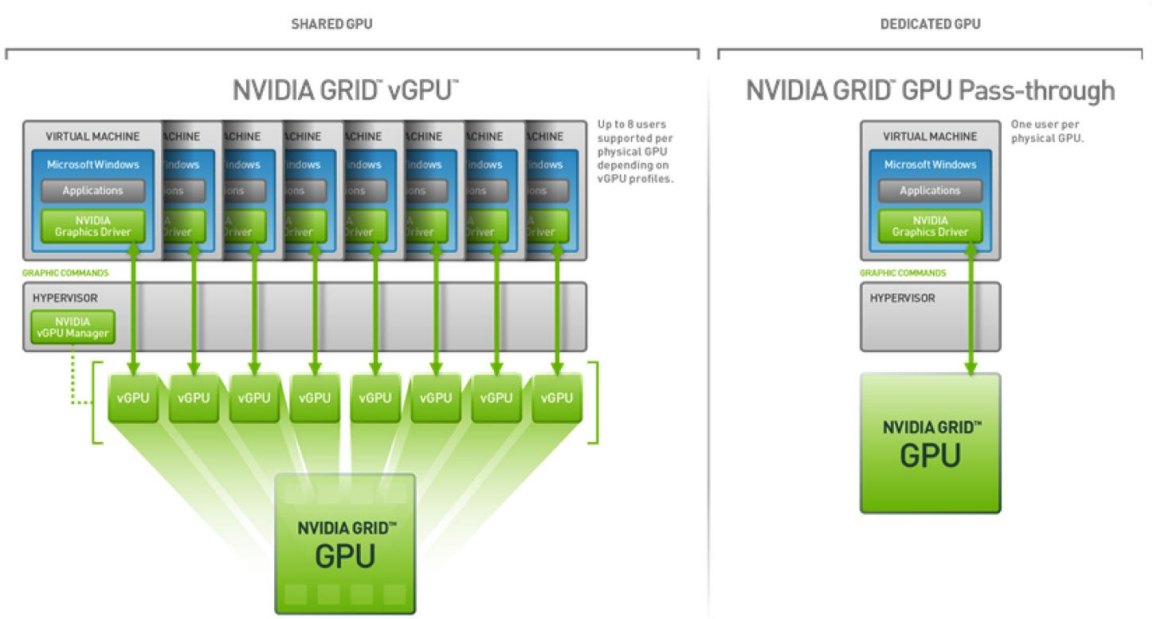
NVIDIA GRID technology offloads graphics processing from the CPU to the GPU in virtualized environments. GRID GPUs are specifically designed for use in virtual environments; the cards lack external video connections. For users, NVIDIA GRID provides a highly responsive experience for demanding 3D graphics applications on any device, even tablets.

NVIDIA GRID supports two modes of operation:

- **GPU pass-through.** Each user is assigned a physical GPU.
- **Virtual GPU (vGPU).** A single physical GPU is shared by up to eight users (NVIDIA K2) for improved user density.

GPU pass-through is well suited to power users such as CAD developers and other 3D application users, and vGPU is optimal for less intensive graphics users. As shown in Figure 3, NVIDIA GRID technology delivers dedicated GPUs for power users as well as for more economical shared GPU options, such as vGPU.

Figure 3) NVIDIA GRID technology (graphic supplied by NVIDIA).



As the first virtualized GPU hardware solution, NVIDIA GRID vGPU allows multiple users to share the graphics processing power of a single GPU. As a result, you can efficiently share GPU resources and broaden the reach of advanced visualization. Each user is assigned a GRID vGPU profile. These profiles correspond closely to the capabilities of typical GPU cards.

Table 1) NVIDIA GRID vGPU profiles match shared GPU users to correct graphics capabilities.

NVIDIA GRID Graphics Board	Virtual GPU Profile	Application Certifications	Graphics Memory (MB)	Max Displays per User	Max Resolution per Display	Max Users per Graphics Board
GRID K2	K260Q	✓	2,048	4	2,560×1,600	4
	K240Q	✓	1,024	2	2,560×1,600	8
	K220Q	✓	512	2	2,560×1,600	16
	K200		256	2	1,900×1,200	16
GRID K1	K140Q	✓	1,024	2	2,560×1,600	16
	K120Q	✓	512	2	2,560×1,600	32
	K100		256	2	1,900×1,200	32

Note: Refer to the NVIDIA website for the latest information.

NVIDIA GRID software includes a complete stack of GPU virtualization, remoting, and session-management libraries that allow users to experience graphics-intensive desktops and applications with exceptional capture, efficient compression, fast streaming, and low-latency display.

The FlexPod Datacenter solution with Citrix and NVIDIA uses the NVIDIA GRID K2, featuring two high-end Kepler GPUs with a total of 3,072 CUDA cores and 8GB of GDDR5 memory. Because each Cisco UCS C240 M4 server supports two GRID K2 devices (4 GPUs), a single 2U server is capable of supporting up to four users with GPU pass-through or up to 32 users with vGPU.

3.5 Citrix XenDesktop with HDX 3D Pro

Citrix XenDesktop with HDX 3D Pro and NVIDIA GRID are integrated with the proven FlexPod architecture to provide a complete desktop virtualization solution. Citrix leads the industry with approximately 3 billion deployed desktops and support for a wide variety of client types. Citrix XenDesktop delivers applications and desktops as secure mobile services to improve mobility and provide greater security for intellectual property with centralized control. XenDesktop with HDX 3D Pro delivers a native, touch-enabled experience that is optimized for the type of device, as well as the network.

HDX 3D Pro technologies are specifically designed to enhance visual performance and deliver high-performance, graphics-intensive applications with multiple WAN optimization technologies, deep compression and quality of service (QoS) controls, hardware-level GPU acceleration, and full support for OpenGL applications.

Standard VDI environments utilize CPU resources to perform graphics functions. XenDesktop was among the first to support the use of GPUs for 3D graphics in virtualized environments and the first hypervisor to support NVIDIA GRID in production. HDX 3D Pro supports pass-through operations as well as vGPU. XenDesktop currently supports Windows® virtual machines, with a technology preview under way for Linux® virtual machines.

By complementing XenDesktop HDX 3D Pro with integrated GRID software and NVIDIA GPUs, FlexPod empowers users everywhere with full graphics capabilities for an optimal end-user experience.

Table 2) FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID solution components.

Component	Functions
NetApp FAS storage	Provides the flexibility to support VDI infrastructure and data storage as needed with proven storage efficiency that reduces storage requirements and simplifies management.
Cisco UCS compute	Unites computing, networking, storage connectivity, and virtualization in a single cohesive system with a compact 2U form factor per server. Service profiles simplify deployment and facilitate failover.
Cisco Nexus switches	Proven network architecture delivers high bandwidth with low latency and superior configuration flexibility.
Citrix XenDesktop HDX 3D Pro	Technologies that optimize the virtualization of 3D graphics, including WAN optimization, deep compression, QoS controls, GPU acceleration, and OpenGL support.
NVIDIA GRID K2 GPUs	Designed for use in virtual environments, NVIDIA GRID supports both pass-through—offering power users an entire Kepler GPU—and the ability to share a single GPU among multiple users.

3.6 Overall Solution Benefits

FlexPod Datacenter with Citrix and NVIDIA delivers a number of benefits to the data center, including:

- **Reduced time to insight.** Centralize important datasets in one or a few locations and make the results of visualization available wherever they're needed. This simplifies data integration between systems and accelerates workflows across functions. Your end users and business decision makers can see important results in near real time, without the bottlenecks that result from transferring huge datasets over network connections or by mail.

  Collaborators view and manipulate the same images, eliminating potential points of confusion and miscommunication and saving valuable time. End users of all types have better access to information for deeper insight and faster decision making.
- **Simplified operations and lower costs.** Eliminate the expense and complexity of dedicated workstations by replacing them with a much more efficient and scalable shared resource capable of supporting workers wherever they happen to be. This solution delivers higher resource utilization for greater return on investment and can reduce the total number of expensive software licenses required. Because all components reside in the data center, this solution is much more reliable than workstations in dispersed locations, where power loss and other events can interrupt work and cause data loss.
- **Ability to deploy infrastructure more quickly.** Whether it's in an existing data center or a remote location, the integrated and tested design of FlexPod Datacenter with Citrix and NVIDIA enables you to have new infrastructure up and running in less time with less effort.

4 Solution Sizing and Getting Started

Because FlexPod Datacenter is well established, many resources exist to facilitate and streamline deployment. The combination of FlexPod and XenDesktop is well understood. A recent NetApp Community article explains the benefits of deploying the two together.

One of the great advantages of the FlexPod architecture over other integrated infrastructure solutions is that all solution elements can be flexibly sized to meet your specific requirements. You can easily tailor your compute and GPU resources, storage, and network bandwidth to satisfy requirements. For information about how to deploy virtual desktops that require high performance, low latency, and a small data center footprint, refer to TR-4342: NetApp All-Flash FAS Solution for Persistent and Nonpersistent Desktops with Citrix XenDesktop and XenApp.

4.1 Sizing Cisco UCS Servers with NVIDIA GRID K2 Cards

You can size your Cisco UCS servers to address the needs of power users who are using GPU pass-through or shared GPU seats through vGPU.

For GPU pass-through, each 2U C240 M4 server supports two NVIDIA GRID K2 cards for a total of four GPUs. As a result, each server can support four pass-through users. The key to sizing servers is simply a matter of appropriately sizing the CPU resources and memory for the most demanding application you need to support. For instance, to support four seats for a particular application, refer to the recommended amount of CPU and memory for a single instance and multiply by four.

The process of sizing for vGPU seats is similar. First, decide which vGPU profile you need to support your desired applications (refer to Table 1). This determines the number of seats you can support per server. Then, given the number of seats, size the server to provide the right number of cores and memory.

4.2 Sizing NetApp FAS Storage

When it comes to sizing NetApp FAS storage for the FlexPod Datacenter with Citrix and NVIDIA solution, two scenarios are possible:

- FAS storage is used for VDI only. Datasets and application data are stored on external storage.
- FAS storage is used for VDI plus dataset and application data storage.

When sizing for VDI only, a NetApp all-flash FAS is a good choice to help achieve high performance with ultra-low latency, storage efficiency, and advanced data management. For example, the FAS8020 model is capable of scaling to deliver up to 50,000 IOPS using an 80% write/20% read workload typical of VDI environments. Contact your NetApp systems engineer for assistance with sizing your environment.

To support both VDI and data storage, the preferred configuration is a four-node (or larger) cluster that includes:

- Two all-flash FAS nodes appropriately sized to support VDI needs as described earlier.
- Two or more hybrid FAS nodes to supply the necessary storage capacity for datasets and application data that require a lot of capacity. NetApp recommends a combination of high-capacity disk drives for storing large visualization datasets with separate high-performance SAS drives for storing application data for applications with high capacity requirements. NetApp Flash Pool™ intelligent caching is configured in each controller to accelerate random read operations.

Alternatively, a single NetApp FAS HA pair can be configured with an all-flash aggregate to support VDI and separate data aggregates to store datasets and application data. Contact your NetApp systems engineer for sizing assistance.

4.3 Estimating Network Bandwidth Requirements

For remote users, NetApp recommends estimating how much bandwidth is required to deliver good performance. Based on guidance in the H.264 Primer, you can estimate the bit rate needed using the following formula:

```
Bit rate = horizontal × vertical × color depth × FPS × motion rank × 0.07
```

Where:

- Horizontal is the horizontal pixel count for desired screen.
- Vertical is the vertical pixel count for desired screen.
- Color depth is the number of bits per pixel.
- FPS is the desired frames per second.
- Motion rank is an estimate of how quickly the screen is changing:
  - 1 means 25% of the screen is constantly changing.
  - 2 means 50% of the screen is constantly changing.
  - 3 means 75% of the screen is constantly changing.
  - 4 means 100% of the screen is constantly changing.
- 0.07 is the H.264 calculation coefficient (codec dependent).

For example, for a 2,560 × 1,600 screen at 24 fps (acceptable quality), color depth of 24, and a motion rank of 2 (which is generous for graphics-intensive applications because much of the screen can be occupied by menus and toolbars that are static):

```
Bit rate = 2560 × 1600 × 24 × 2 × 0.07 = 13.7Mbps
```

That's approximately 1% of a 1Gbps link. For the sake of comparison, transferring a 500GB file over a 1Gbps link takes over an hour.

Note: It's not uncommon for LTE cellular networks to deliver this amount of bandwidth.

5 Solution Architecture

To demonstrate the viability of this solution, a sample of the architecture hardware was assembled, and the required software stack was installed to enable a baseline hardware and software platform upon

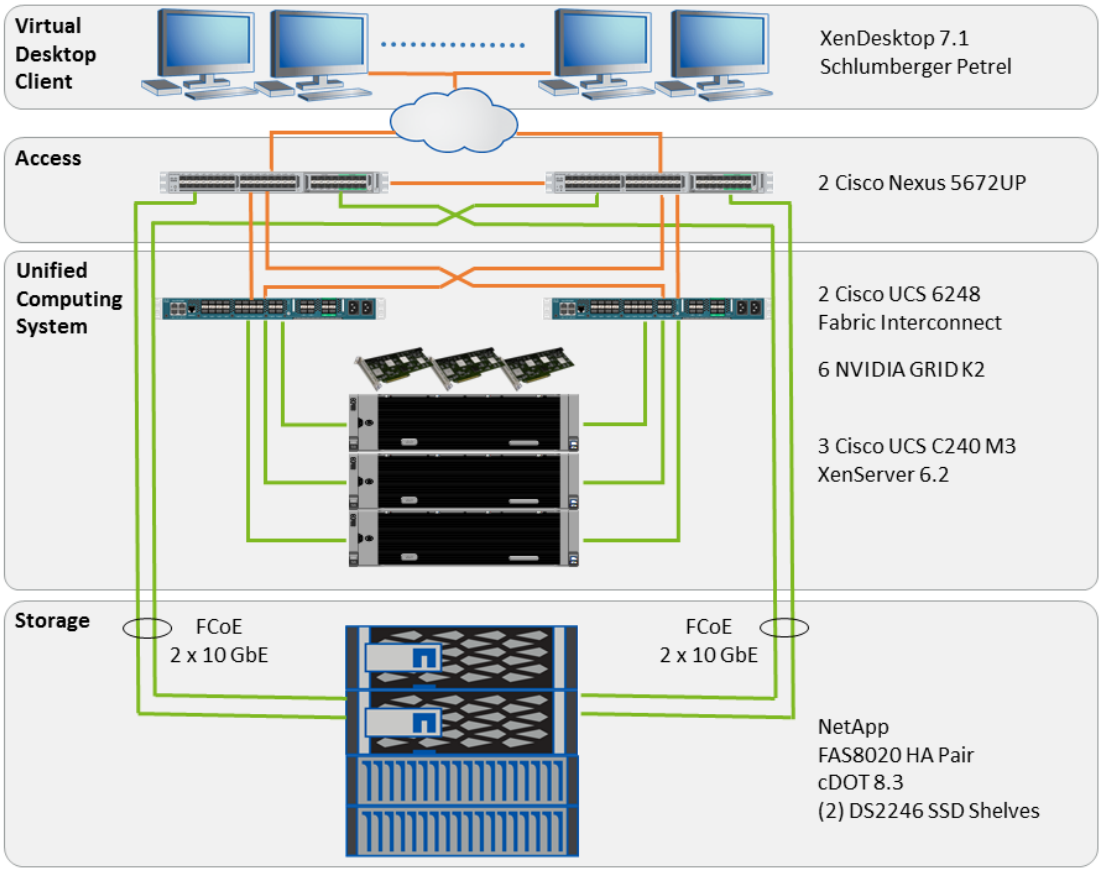
which any of the high-performance applications could be operated. In our study, we used a FlexPod system with XenServer. The following components were used in the build-out:

- 1 NetApp FAS8020 HA pair with 2 DS2246 disk shelves running clustered Data ONTAP 8.3
- 3 Cisco UCS C240 M3
- 6 NVIDIA GRID K2 cards (2 per server)
- 2 Cisco UCS 6248 fabric interconnects
- 2 Cisco Nexus 5672UP switches
- XenServer 6.2
- XenDesktop 7.1

Installation and configuration of the components were patterned after the reference architecture described in TR-4342: NetApp All-Flash FAS Solution for Persistent and Nonpersistent Desktops with Citrix XenDesktop and XenApp.

Figure 4 depicts the end-to-end architecture.

Figure 4) FlexPod XenServer NVIDIA GRID architecture.



The FAS8020 HA pair served the boot images for the Cisco UCS C240 servers by using the FCoE protocol. FAS8020 storage systems stored the images for each of the VDI clients in volumes located on the SSD drives in the DS2246 shelves. A separate volume on the FAS8020 stored application data. Deduplication was enabled for space-efficient storage of the cloned images. Clustered Data ONTAP 8.3 was selected to run on the storage platform.

In this trial, the application tested was Schlumberger Petrel, which enables highly visual 3D modeling and representations of large seismic databases. Using this configuration and employing the NVIDIA GRID cards in a VDI environment, the engineers experienced a productive capability comparable to that of a traditional high-cost desktop workstation environment. The user experience was smooth and rich, with low latency and excellent scalability.

For other high-performance applications, a hardware and software stack similar to the one described could serve as a reference, substituting only the application layer. Sizing may adjust the hardware model of the storage and servers used.

In conclusion, the combination of FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID is a highly capable solution for virtual desktop environments that require accelerated graphics performance.

References

The following resources were referenced in this document:

- DS-3614: FlexPod Datacenter with Citrix XenDesktop and NVIDIA GRID
  www.netapp.com/us/media/ds-3614.pdf
- TR-4342: NetApp All-Flash FAS Solution for Persistent and Nonpersistent Desktops with Citrix XenDesktop and XenApp
  www.netapp.com/us/media/tr-4342.pdf
- WP-7199: Future of Mobility in Exploration and Production
  https://private-communities.netapp.com/docs/DOC-33761
- Next-Generation Data Center Architecture for Advanced Compute and Visualization in Upstream Oil and Gas
  https://private-communities.netapp.com/docs/DOC-30028
- XenDesktop on FlexPod
  http://community.netapp.com/t5/Tech-OnTap-Articles/XenDesktop-on-FlexPod/ta-p/85185
- Cisco UCS C240-M3 Rack Server with NVIDIA GRID GPU Cards on Citrix XenServer 6.2 and XenDesktop 7.5
  www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper_C11-732283.html
- Citrix Joins Cisco and NetApp in Delivering FlexPod Cooperative Support
  www.citrix.com/content/dam/citrix/en_us/documents/partner-documents/citrix-joins-cisco-and-netapp-in-delivering-flexpod-cooperative-support.pdf?accessmode=direct
- FlexPod Datacenter with Citrix XenDesktop 7.2 and VMware vSphere 5.1
  www.cisco.com/c/dam/en/us/td/docs/unified_computing/ucs/UCS_CVDs/cisco_ucs_xd7esxi51_flexpod.pdf
- H.264 for the Rest of Us
  http://wwwimages.adobe.com/content/dam/Adobe/en/devnet/video/articles/h264_primer/h264_primer.pdf
- NVIDIA GRID Enterprise Graphics Virtualization
  www.nvidia.com/object/virtual-gpus.html

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.