HPE Reference Architecture for client virtualization in healthcare on HPE Synergy Composable Infrastructure

Delivering optimal user experiences on demand with Citrix Cloud using NVIDIA T4
Contents

Executive summary .............................................................................................................. 4
Solution overview .................................................................................................................. 5
  HPE Synergy ...................................................................................................................... 11
  HPE Synergy 12000 Frame ............................................................................................... 14
  HPE Synergy Composer 2 ................................................................................................. 15
  HPE Synergy 480 Gen 10 Compute Module ...................................................................... 16
  PCIe graphics expansion module ..................................................................................... 17
  HPE 3PAR 8440 Storage .................................................................................................. 19
  NVIDIA GRID .................................................................................................................. 21
  NVIDIA Licensing ............................................................................................................ 22
  NVIDIA T4 GPU ............................................................................................................. 23
  HPE OneView .................................................................................................................. 24
  Citrix Cloud ..................................................................................................................... 24
  Citrix Virtual Apps and Desktops Service .......................................................................... 25
  Citrix HDX 3D Pro for GPU based medical imaging applications .................................. 26
  Citrix Content Collaboration ......................................................................................... 26
Solution components ........................................................................................................... 27
  Hardware ........................................................................................................................... 27
  Software ........................................................................................................................... 27
Best practices and configuration guidance .......................................................................... 27
  HPE Synergy configuration ............................................................................................. 27
  VMware vSphere Server Configuration ............................................................................ 33
  Citrix Virtual Apps and Desktops Service Configuration .................................................. 34
Capacity and sizing ............................................................................................................. 34
  Login VSI .......................................................................................................................... 34
  Non-Accelerated Office Worker workload ....................................................................... 37
  Accelerated Office Worker workload with NVIDIA T4 GPU .......................................... 39
Summary ............................................................................................................................. 42
Appendix A: Benefits of HPE Synergy Composable Infrastructure in VDI/virtual environments ......................................................................................................................... 43
Appendix B: Bill of materials .............................................................................................. 44
Resources and additional links .......................................................................................... 46
Executive summary

As the healthcare industry grows and becomes hyper-connected, the IT teams that support healthcare organizations must find better ways to leverage technology to improve overall patient care, work experiences for physicians and staff, data confidentiality, and regulatory compliance. Integrating new, future-ready technologies will help to ensure the right experience for each end user (patients, doctors, nursing staff, etc.). The challenge is to deliver this experience cost-effectively, with the right mix of technology.

End user computing has maintained a strong position in the healthcare industry for well over a decade because it addresses many of the challenges listed above. Also, it is important to understand where the power of graphic processing units (GPUs) fits in, to optimize user experience while managing overall costs. Adopting GPU everywhere provides a quality experience, however, not every user needs a GPU. Moreover, deploying GPUs for users that do not need them is expensive. It is critical to understand and address the specific needs of various workers - with and without GPU requirements – in a healthcare setup.

This Reference Architecture demonstrates how to use HPE Synergy, the Citrix® Virtual Apps and Desktops Service, and NVIDIA® to achieve the outlined goals. HPE Synergy offers the flexibility to provision and delivers the right mix of resources – both non-accelerated (non-GPUs) and accelerated virtual apps with GPU. This solution provides the workstations with “as-needed” access to graphical applications but does not require virtualized GPUs (vGPUs) in workstations and graphical acceleration.

Working with Citrix and NVIDIA allows us to present:

- Virtual desktops and applications
- GPU Accelerated applications
- GPU Accelerated desktops

HPE Synergy can uniquely deliver this solution with the highest user density for virtual desktop infrastructure (VDI) and server delivered applications. To reiterate, HPE Synergy delivers more VDI sessions per rack for medical applications and electronic health records (EHR) while maintaining Data Security and providing the best user experience.

Other unique security innovations on HPE Synergy include the Silicon Root of Trust technology that extends across the lifecycle for end-to-end security. HPE 3PAR enables hosting user data and hosting VMs in the form of the solution stack and allows scaling.

This Reference Architecture defines the best cost-effective solution for healthcare organizations that can deploy to deliver the right experience for their end users. This document demonstrates the following:

- Benefits of running accelerated GPU workloads in EUC environment leveraging the strength of HPE Composable Infrastructure, Citrix, and HPE 3PAR
- High-performance GPUs to accelerate the applications with an “as-needed” consumption model
- GPUs can be repurposed from display mode to compute mode, for addressing both visualization and analytics workload
- Composable infrastructure provides flexibility for diverse and demanding healthcare workloads on a single platform

Target audience: This document provides recommendations for configuring EUC solutions for healthcare with HPE Synergy platform. This will educate IT decision makers, enterprise architects, and partners and provide understanding about the groundbreaking HPE Composable Infrastructure, which enables you to compose fluid pools of physical and virtual compute, storage, and fabric resources into any configuration for diverse workloads.
Solution overview

This Reference Architecture highlights the value of HPE Synergy Composable Infrastructure, GPUs, and the Virtual Apps and Desktops Service in Citrix Cloud. It defines how this combination delivers a well-designed client virtualization solution for healthcare organizations.

High-end graphics applications and GPU enhanced artificial intelligence and inferencing applications continue to change the nature of healthcare. Also, video and other graphics technologies need graphics-acceleration, making GPUs a core component for a healthcare solution. On the other hand, practitioners spend most of their time on text-based electronic health records (EHR) which does not need high-performance graphics acceleration. Hence this solution is designed to optimally address computing needs of both text and graphics workloads.

Figure 1 describes the healthcare system use cases with typical percentage distribution.

![Healthcare System Use Cases](image)

**Figure 1.** Healthcare system use cases (Source: Independent Research)

Architecting a right-sized VDI solution for healthcare includes understanding the usage pattern and workload types. As shown in Figure 1, this has only a limited number of users who would use accelerated applications and hence this solution is designed to virtualize the high-end graphics applications instead of running them on individual user workstations. In this way, high-performance GPUs can be deployed where the application will use them most effectively. Also, GPU resources are used on-demand instead of being pre-allocated to individual desktops. To summarize, application virtualization provides access to high-end GPUs at a lower cost due to the efficiency of an on-demand model and these lower costs are translated to lower healthcare cost per capita.
Figure 2 illustrates the healthcare system architecture. This outlines the delivery of non-graphical and graphical applications across a healthcare organization, which includes EHR traffic, accelerated 3D graphics traffic, and DICOM/H7 traffic.

This Reference Architecture provides a framework for client virtualization to deliver a quality solution, which enhances end user experience, improves performance, and provides increased user density while improving security and manageability.
Figure 3 highlights the solution components and provides an architectural overview of the client virtualization solution. This solution is designed with nine HPE Synergy 480 Gen10 Compute Modules. The first four modules feature four NVIDIA T4 GPUs, each module provides enhanced visualization for published applications-desktops and inferencing. The remaining five modules do not feature GPUs, which caters to the majority of healthcare workers including those that interact with EHR systems. Virtual machines, EHR data, and other healthcare centric information are hosted on HPE 3PAR 8440 StoreServ Storage. Citrix Content Collaboration enables collaboration between healthcare personnel.

**Figure 3. Solution architecture overview**
HPE Synergy Composable Infrastructure
HPE Synergy Composable Infrastructure is powered by the HPE Synergy Composer appliance and managed by HPE OneView. HPE Synergy can compose and decompose compute, storage, and network to the respective servers on-demand. The solution consists of the HPE Synergy 480 Gen10 Compute Module with PCIe expansion module hosting NVIDIA GPUs to facilitate the Citrix app and desktop workloads.

HPE 3PAR 8440 unified All-Flash Array
HPE 3PAR 8440 unified All-Flash Array provides a unified storage solution to host the virtual desktops and file shares. HPE 3PAR All-Flash Array provides accelerated read and write performance and low latency, thus providing great application response time at peak usage and boot storms. The solution is designed aptly to host the healthcare medical data, diagnostic reports, high-quality medical images and EHR database on HPE 3PAR 8440 unified All-Flash Array as file data is managed and administered using HPE 3PAR File Persona. Since the file data resides on All-Flash disks, the read/write data provides excellent response time to end users.

NVIDIA T4 GPU
Physicians depend heavily on accessing the diagnostic medical scan reports of patients and healthcare consumers. The solution has been designed precisely to address the challenges of user experience by healthcare personnel. HPE Synergy with NVIDIA T4 GPU provisions vGPU to Citrix Virtual Apps and Desktops, as these accelerated applications and desktops are accessed by medical staff who gets benefit from precise and detailed views of the medical images. This platform is backed by the NVIDIA GPU and Citrix HDX 3D Pro technology thereby providing doctors & medical staff a high-quality image.

Citrix Cloud
Citrix Cloud is a cloud-based control plane of service offerings. It includes services such as Citrix Virtual Apps and Desktops Service and Content Collaboration. The design contains a demarcation of Citrix infrastructure components residing in the Citrix Cloud management plane and user workloads running within the data center. The Citrix Cloud Connector is the component that enables the connectivity between resource locations hosting the Citrix Virtual Desktop Infrastructure on HPE Synergy and the Citrix Cloud control plane. It is installed on two or more Windows servers per resource location for redundancy and performance.

Citrix Content Collaboration
Citrix Content Collaboration enables end users to access collaborative medical data and content securely from anywhere. This allows doctors and healthcare staff to access content within the organization. At the same time, it also enables outpatient users, remote healthcare personnel, and remote doctors to access the same data anywhere and from any device.

Citrix Virtual Apps and Desktops Service
End users have Citrix Virtual Apps and Desktops provisioned on HPE Synergy and access their respective desktops, medical data, and applications via endpoints. Citrix Machine Creation Services (MCS) is used to provision accelerated and non-accelerated virtual machines for medical and administrative staff to help them perform their day-to-day activities.
Figure 4 illustrates the interaction between Citrix Cloud, HPE Composable Infrastructure, HPE 3PAR unified Storage hosting the Virtual Desktop Infrastructure and the end users.

For architecture planning, workers can be categorized into different types. The three different user segments and their entitlements are:

- Healthcare users (GPU-Shared) have a shared NVIDIA GPU infrastructure on HPE Synergy from which accelerated virtual apps and desktops are delivered to medical staff, such as doctors and nurses.

- Administrative staff, out-patients, and mobile doctors (standard workload) have access to resources that do not typically require graphical acceleration. This user group will be utilizing productivity apps like Microsoft® Office, records management software, and other non-

- Power users and operators who have a dedicated NVIDIA vGPU profile aligned to the virtual desktop for special use cases like image inferencing software, deep learning, and professional 3D applications, etc.

For all three user segments, the access mechanism is via Citrix Workspace app.
Figure 5 highlights three different user segments and their entitlements.

Figure 5. User segments and their workload types
At the core of this solution are the underlying building blocks. These blocks lay the foundation of the digital workspace solution and configured together in different ways to solve unique customer needs. In Figure 6, the HPE Synergy 12000 frame is hosting nine (9) HPE Synergy 480 Gen10 Compute Module and twelve (12) NVIDIA T4 cards for catering computational and graphical needs of the workload. The HPE 3PAR StoreServ 8440 AF Storage Array provides the speed and storage capacity for our solution.

The key building blocks of this solution are:

- HPE Synergy 480 Gen10 Compute Modules
- HPE PCIe Expansion Module with NVIDIA T4 Cards
- HPE 3PAR StoreServ 8440 AF Storage Array

**Figure 6.** HPE Synergy, HPE 3PAR, and NVIDIA components used in this solution

**HPE Synergy**

HPE Synergy, the first platform built from the ground up for composable, offers an experience that empowers IT to create and deliver new value instantly and continuously. A single infrastructure reduces operational complexity for traditional workloads and increases operational velocity for new breeds of applications and services. HPE Synergy empowers IT administrators and developers, to use infrastructure-as-code for deploying and managing their data center environments. This new approach for composable infrastructure combines true stateless computing with rapid deployment and updates.
Figure 7 highlights the infrastructure evolution from silo to composable.

**Fluid resource pools**

HPE Synergy allows the transformation of traditionally rigid physical systems into flexible virtual resource pools. HPE Synergy creates a resource pool of “stateless” compute, storage, and fabric capacity that can be configured almost instantly to rapidly provision infrastructure for a broad range of applications. Now IT can manage both infrastructure growth and shrinkage, from a single fluid pool of resources not stranded in static silos. The smaller infrastructure reduces Capital Expense (CapEx), and better management of provisioning reduces power and cooling consumption, minimizing operating expense (OpEx). Fluid resource pools are easily expandable by automatically integrating and assembling additional infrastructure readily composed (and re-composable) to meet the changing workload demands of the business. Figure 8 shows how HPE Synergy fluid resource pools help in optimal utilization.
Software-defined intelligence

The software-defined intelligence in HPE Synergy reduces operational complexity and enables IT organizations to make needed programmatic changes quickly and confidently with minimal human intervention. HPE Synergy abstracts operational details and replaces them with high-level, automated operations. HPE Synergy uses templates to enable automatic implementation for change operations such as updating firmware, adding additional storage to a service, or modifying the network. Figure 9 illustrates how HPE Synergy abstracts away operational complexities with the help of software-defined intelligence.

Figure 9. Software-defined intelligence- comparison of traditional versus composable Infrastructure

Unified API

Traditional IT management typically involves low-level API abstractions with a number of orchestration applications. Every device has its own API and each API has error code formats. Using multiple interfaces requires knowledge of all APIs and which is complex and time-consuming. The task to automate processes across heterogeneous interfaces is complicated since those APIs tend to be at a very low (CLI) level that requires configuration of each component separately. Getting a single server up and running can require as many as 500 or more individual calls to low-level tools to get the infrastructure configured properly. HPE Synergy includes a high-level unified API that brings together all the resources (compute, storage, and fabric) under a single interface with a single data format. Abstracting the API to a high level simplifies programmability.
HPE Synergy 12000 Frame

The HPE Synergy 12000 frame is a key element that provides the base for an intelligent infrastructure with embedded management and scalable links for expansion. The HPE Synergy 12000 frame is the base infrastructure that pools resources of compute, storage, fabric, cooling, power, and scalability. With an embedded management solution combining the HPE Synergy Composer and HPE Synergy frame Link Modules, IT can manage, assemble, and scale resources on demand. The HPE Synergy 12000 frame is designed for current and upcoming needs with expanded compute and fabric bandwidths.

Figure 10. HPE Synergy 12000 frame
HPE Synergy Composer 2
HPE Synergy Composer 2 provides enterprise-level management to compose and deploy system resources for your applications. This management appliance uses software-defined intelligence with embedded HPE OneView to aggregate compute, storage and fabric resources in a manner that scales to your application needs, instead of being restricted to the fixed ratios of traditional resource offerings. HPE OneView server profiles and profile templates capture the entire server configuration in one place, enabling administrators to replicate new server profiles and to modify them as needed to reflect changes in the data center. With HPE OneView Rest API and automation tools, the entire process of server personality definition and configuration can be automated. In this solution, the HPE OneView REST API and PowerShell library are used to automate the server profile application to “stateless” servers.

Figure 11. Front view of HPE Synergy Composer node

What’s New?
HPE Synergy Composer 2 comes with Gen10 architecture and equipped with UEFI, high assurance boot, iLO5 and two additional 10GB network ports. The memory is increased from 16GB to 64GB and storage is increased from 240GB SATA to 400GB NVMe. The processing power has also been increased from 2-core / 4-threads @ 2.4GHz to 8-core / 16-threads @ 2.0GHz.
HPE Synergy 480 Gen 10 Compute Module
The HPE Synergy 480 Gen10 Compute Module delivers superior capacity, efficiency, and flexibility in a two-socket, half-height form factor to support demanding workloads. This module provides a composable compute resource that is auto-discovered, inventoried, quickly provisioned, easily managed, and seamlessly redeployed to deliver the right compute capacity for changing workload needs.

Figure 12. Front View of HPE Synergy 480 Gen10 Compute Module

Freed of stranded compute resources, enterprise data centers can now deliver the right compute capacity for changing workload needs. HPE Synergy computes superior, enterprise-grade availability and offers quick and confident infrastructure changes. The change operations such as firmware updates can be applied instantly for initial set up or staged so that they can automatically have the effect later.

Benefits of HPE Synergy Composable Infrastructure for end-user computing environments
HPE Synergy, a powerful software-defined solution, enables you to compose fluid pools of physical and virtual compute, storage, and fabric resources into any configuration. The benefits are to gain efficiency and control, and deploy IT resources quickly through a single interface.

- HPE Synergy Composable Infrastructure can be composed targeting different VDI workloads with respect to GPU and non-GPU workloads. During different time-intervals when the consumption of GPU declines, HPE Synergy Compute Module can be recomposed with a non-GPU server profile for non-accelerated workloads or decompose HPE Synergy resources thereby using the infrastructure optimally.

- HPE Synergy Compute Modules in the two-slot form factor comprises of expansion module that can host a large number of GPUs when compared to a single form factor compute module. This solution has HPE Synergy with PCIe expansion module, which is capable of hosting 4x NVIDIA T4 GPUs to address the graphics intensive Citrix Virtual Apps and Desktops.

- HPE Synergy Composable Infrastructure can be composed and decomposed based on the consumption of desktops by end users in the VDI environment thereby optimally consuming the power in the data center. For instance, VMware vSphere Distributed Power Management can be enabled on the vSphere clusters that could trigger ESXi hosts to go into standby mode when there is a decline in the consumption of compute resources further enabling the data center administrator to decompose HPE Synergy resources. The result is reduction of power consumption and an increase in green footprint across the data center.

- HPE Synergy's composability features like compose, re-compose, and decompose can be achieved via Unified API in conjunction with VDI desktop pool provisioning using different automation methodologies. Tools and programming languages such as vSphere PowerCLI, Microsoft PowerShell and Python can be used to orchestrate and provision HPE Synergy Fluid Resource Pools comprising of compute, storage, and network via Unified API as and when required for VDI Infrastructure. VDI desktop pools can be automated to be provisioned in a timely fashion to consume the readily available composable infrastructure, thus bringing in an efficient and optimized consumption methodology.

- HPE Synergy Composer can be leveraged to deliver automation through a unified API that provides a single interface to discover, inventory, configure, provision, update, and diagnose the composable infrastructure in a heterogeneous environment. This fully programmable interface integrates into dozens of popular management tools such as Microsoft System Centre, VMware vCenter® and DevOps tools such as Chef and Ansible thus making HPE Synergy Composable Infrastructure an ideal platform for virtual environments. This solution uses simpler and widely used tools, such as Windows PowerShell and VMware PowerCLI to demonstrate the on-demand composable feature of the HPE Synergy platform.
**PCIe graphics expansion module**

Hewlett Packard Enterprise offers graphics accelerator options for HPE Synergy 480 Gen10 Compute Modules to host multiple graphics cards in MXM format or 2 Full-Length, full height, double-wide configuration. The expansion modules are connected to the HPE Synergy 480 Gen10 Compute Modules through PCIe pass-through mezzanine cards located in the compute module. This flexible solution supports the spectrum of graphics cards like NVIDIA Quadro P6000, RTX 6000 and NVIDIA Tesla® P40, P6, M10 & T4 to meet the customer requirements by providing high consolidation ratio for accelerated workloads. This particular solution uses NVIDIA T4 cards with PCIe expansion for graphics acceleration of virtual app and desktop workloads.

*Figure 13. Cross sectional view of a PCIe graphics expansion module*

*Figure 14 demonstrates composable capabilities of HPE Synergy platform.*
Figure 15 depicts how HPE Synergy include a high-level, unified API that brings together all the resources, compute, storage, and fabric under a single interface.

**Figure 15.** Comparison of programming interface - Traditional versus composable infrastructure
HPE 3PAR 8440 Storage
The amount of data generated and processed in today’s healthcare environments is massive. The storage solution used should be robust enough to not only support but to accelerate healthcare by enabling organizations to provide personalized care by understanding their needs, reducing risks and expediting decision making. HPE 3PAR StoreServ 8440 Storage ensures high throughput and consistently low latency even with demanding workloads of EHR databases, Unstructured DICOM / PACS images along with highly consolidated VDI workload running medical and productivity applications. Every healthcare organization wants to deliver an exceptional experience and with the advent of software-defined workspace, the storage plays a vital role. One of the most important success criteria is to ensure that the Citrix Virtual Apps and Desktops are as good as their traditional systems (or preferably better). The HPE 3PAR StoreServ 8440 Storage platform is an enterprise-class flash array that helps you consolidate primary storage workloads (file, block, and object) offering flexible I/O host connectivity without compromising performance, scalability, data services, or resiliency. The purpose of this new HPE 3PAR model based on the proven HPE 3PAR architecture is built for the all-flash consolidation, delivering the performance, simplicity, and agility needed to support your hybrid IT environment. HPE 3PAR StoreServ 8440 Storage is available in a single all-flash model, the 8440, that offers rich Tier-1 data services, quad-node resiliency, fine-grained Quality of Service (QoS), seamless data mobility between systems, high availability through a complete set of persistent technologies, and simple and efficient data protection with a flat backup to HPE StoreOnce backup appliances.

Figure 16. HPE 3PAR StoreServ 8440 Storage

The HPE 3PAR StoreServ 8440 Storage is designed for true convergence of block, file, and object access to help consolidate diverse workloads efficiently. In this solution, the EHR database resides on all-flash storage with assured QoS to accelerate EHR applications.
Figure 17 depicts how the HPE 3PAR file persona provides file storage for high definition medical images on an NFS share via Citrix Content Collaboration to mobilize the health data securely across devices. It is also hosting a CIFS share for storing user profile data to ensure user settings are applied to the user’s virtual desktop and applications, regardless of the location and endpoint device for a seamless experience.
**NVIDIA GRID**

NVIDIA GRID architecture enables NVIDIA GPUs to power virtual desktops and applications in a data center environment and accelerate workflows resulting in optimized user experience and improved productivity. Figure 18 depicts how the NVIDIA GRID software works in conjunction with the hypervisor and virtualizes the GPU cards so that they can be shared across multiple virtual desktops or applications, while the guest OS on the VM accesses the vGPU like a pass-through device.

![NVIDIA GRID Architecture Diagram](image)

*Figure 18. NVIDIA GRID architecture*
Table 1 shows the detailed list of NVIDIA GRID components and their function.

Table 1. List of GRID architecture components:

<table>
<thead>
<tr>
<th>Components</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA T4 GPU</td>
<td>NVIDIA T4 is a graphical processing unit based on latest Turing™ architecture for data centers</td>
</tr>
<tr>
<td>Virtual GPU Manager</td>
<td>This software runs in the hypervisor layer and allocates physical GPU resources to VMs.</td>
</tr>
<tr>
<td>NVIDIA driver</td>
<td>Driver software facilitates direct GPU access to the VMs.</td>
</tr>
<tr>
<td>Para virtualized interface</td>
<td>Interface for non-performant management operations.</td>
</tr>
<tr>
<td>Framebuffer</td>
<td>Frame buffer or graphics memory is a dedicated resource in NVIDIA GRID for driving video display.</td>
</tr>
<tr>
<td>License Server</td>
<td>License manager for NVIDIA GRID.</td>
</tr>
</tbody>
</table>

**NVIDIA Licensing**

High-end graphics applications and GPU-driven artificial intelligence continue to change the nature of healthcare but at the same time, most of these accelerated applications are used for a limited period of time and assigning a dedicated GPU for these types of workload results in underutilized resources and increased licensing and ownership costs. This solution highlights the benefits of accelerating applications rather than the desktops to optimize the utilization of GPU resources, reduced licensing costs and a green data center. Nevertheless, there are some use cases where the users need high-end graphical computing power throughout their workday and in those cases, assigning dedicated vGPU makes more sense.

Accelerated virtual apps use the GRID vApps licensing model that addresses the requirements for most of the graphic-intensive workloads in a typical healthcare environment, which includes viewing DICOM images, running PACS application and viewing electronic medical records (EMR). Alternatively, if a user needs to run a professional 3D application on an accelerated virtual desktop, a GRID vPC license can be assigned to NVIDIA accelerators and can be licensed according to the workload as illustrated in Figure 19. GRID vPC, GRID vApps, and Quadro vDWS are available on a per Concurrent User (CCU) model. NVIDIA vGPU editions can be purchased by enterprises either as perpetual licenses or as an annual subscription.

![Figure 19. NVIDIA GRID Licensing models and use cases](image)
A perpetual enterprise license allows the hardware to be used indefinitely. Users opt to license using this model are required to subscribe to SUMS for three years. The Support Updates and Maintenance Subscription (SUMS) can be renewed on a yearly basis after the expiration of the initial subscription. Alternatively, NVIDIA GPUs can be licensed with an annual enterprise subscription, which is active for a fixed period as defined by the terms of the subscription license. To be kept active, the license will need to be renewed at the end of the subscription period. The subscription license includes the software license and production level SUMS for the duration of the license subscription period.

NVIDIA GRID licensing model supports a broad spectrum of vGPU profiles that helps in aligning resources aptly and avoids under/over-provisioning of vGPUs. The servers running on accelerated applications can be configured with 16GB profile, which will be accessible by multiple users. Additionally, NVIDIA offers profile sizes of 1GB, 2GB, 4GB, 8GB, and 16GB for Quadro vDWS. Similarly, the graphics workload can be accelerated with 1GB or 2GB profiles for the accelerated desktops.

**Note**

### NVIDIA T4 GPU

NVIDIA T4 GPUs are the latest breed of data center graphics accelerators based on Turing architecture, which enables the acceleration of mixed workloads like virtual machines (VMs) and deep learning, AI, inferencing, high-performance computing (HPC), and other applications. With HPE Composable Infrastructure and NVIDIA GPU dual-mode capability, we can run different workloads.

Table 2 provides information about NVIDIA T4 specifications.

<table>
<thead>
<tr>
<th>Specifications</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Architecture</td>
<td>NVIDIA Turing</td>
</tr>
<tr>
<td>Turing Tensor Cores</td>
<td>320</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>2560</td>
</tr>
<tr>
<td>Memory</td>
<td>16GB GDDR6</td>
</tr>
<tr>
<td>Memory bandwidth</td>
<td>320 GB/s</td>
</tr>
<tr>
<td>Peak single precision floating point performance</td>
<td>8.1 TFlops</td>
</tr>
<tr>
<td>Peak double precision floating point performance</td>
<td>254.4 GFlops</td>
</tr>
</tbody>
</table>

NVIDIA T4 GPUs with their low profile, single-slot form factor are extremely powerful and versatile, which makes them ideal for data center based deployment and with NVIDIA's Compute Unified Device Architecture (CUDA), the programs get direct access to parallel computing elements of the GPU which allows enterprises to run accelerated workloads.

Figure 20 depicts NVIDIA T4 GPU.
Note
NVIDIA Encode (NVENC) technology with Citrix HDX 3D Pro offloads the encoding of both H.264 and H.265 video output from the server CPU and move this workload to the GPU resulting in faster processing and improved performance of hosted applications while saving CPU cycles.

Refer to NVIDIA Grid deployment guide for Citrix XenDesktop on VMware vSphere at http://images.nvidia.com/content/pdf/grid/guides/NVIDIA_GRID_Citrix_XenDesktop_Deployment_Guide.pdf

HPE OneView
HPE OneView is an infrastructure automation engine that simplifies operations to increase the speed of IT delivery for new applications and services. Through software-defined intelligence, HPE OneView brings a new level of automation to infrastructure management by taking a template-driven approach to provisioning, updating, and integrating compute, storage, and networking infrastructure. It is designed with a modern, standard-based application-programming interface (API) and supported by a large and growing partner ecosystem. HPE OneView also makes it easy to integrate powerful infrastructure automation into existing IT tools and processes.

Citrix Cloud
Citrix Cloud contains cloud-based services that enable modern digital workspace with networking and analytics. It’s Citrix Virtual Apps and Desktops Service helps an organization in offloading the core components of the delivery infrastructure to the cloud, where Citrix manages the installation, maintenance, and upgrade of components while IT can focus on managing applications, desktops, policies, and user access.

The Citrix Virtual Apps and Desktops Service contain the following components fully-managed in the cloud by Citrix Delivery Controllers, Citrix Studio, Citrix Director, SQL Servers, the license server, and optionally StoreFront and Gateway. The StoreFront and Gateway are collectively referred to as the Control Layer. The Resource, Platform, Operations, and User Layers are managed within the enterprise, the Access Layer is comprised of the StoreFront, and the Gateway. This layer is managed by Citrix, or you can have those components on-premises.

Figure 21 highlights different Citrix Cloud layers and their components for a better understanding.
Citrix Cloud layers are as follows:

- **User Layer** – This layer encompasses users and their devices.
- **Access Layer** – This layer defines how a user group gains access to their resources while providing secure access policies and desktop/application stores.
- **Control Layer** – This layer represents the Citrix management layer and supports user access to resources. The delivery controllers authenticate users and enumerate resources from StoreFront while creating, managing, and maintaining the virtual resources. All configuration information about the Citrix Virtual Apps and Desktop environment is stored within a back-end cloud-hosted database and the licensing server manages the licensing. With the Citrix Cloud model, you create, deploy, and manage apps and desktops for your users on your preferred platform without the overhead of installation, setup, configuration, maintenance, and monitoring of Citrix Control Layer Infrastructure. By choosing this model, an enterprise can focus on user applications and desktops and offload the deployment, maintenance, and upgrades of core components to Citrix.
- **Resource Layer** – This layer is the focal point of the solution as this layer encompasses Citrix Virtual Apps and Desktops, which will be accessed by the end users.
- **Platform Layer** – This layer comprises infrastructure components like compute, storage, and network fabric, hypervisor, virtual environment manager, and this layer hosts resource layer.
- **Operations Layer** – This layer contains the procedures and tools that support the core product and solution.

**Citrix Virtual Apps and Desktops Service**

The Citrix Virtual Apps and Desktops Service provide a cloud-based management plane, which helps in the secure remote delivery of applications and desktops with centralized management. This service helps end users to access their resources from a plethora of devices like PCs, MacBook’s, thin clients, zero clients, tablets, and smartphones.

Citrix Virtual Apps and Desktops Service addresses multiple challenges faced by healthcare providers. The challenges are:

- Enables secure and quick access to medical applications and data
- Improves compliance and security
- Reduces cost and complexity for administrators
- Increases uptime and availability

Citrix Virtual Apps and Desktops Service manages delivery by utilizing the following components as shown in Table 3.

<table>
<thead>
<tr>
<th>Component</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDX 3D Pro Technology</td>
<td>Remote display protocol that delivers a high-definition 3D user experience of virtual desktops and applications by utilizing graphics processing unit (GPU) for hardware acceleration.</td>
</tr>
<tr>
<td>Thinwire</td>
<td>Thinwire is the default display remoting technology used in Citrix Virtual Apps and Desktops Service.</td>
</tr>
<tr>
<td>StoreFront</td>
<td>Provides the user interface for accessing applications and desktops with self-service capabilities.</td>
</tr>
<tr>
<td>Controller (Citrix Cloud)</td>
<td>User access management via user or computer-based policies.</td>
</tr>
<tr>
<td>Citrix Studio (Citrix Cloud)</td>
<td>Centralized portal for infrastructure and resource management of application and desktops delivery.</td>
</tr>
<tr>
<td>Citrix Director (Citrix Cloud)</td>
<td>Monitoring hub for helpdesk staff to get diagnostic information about users, applications and desktops.</td>
</tr>
<tr>
<td>Citrix Workspace app</td>
<td>Universal client software running on the physical endpoint to access remote applications and desktops.</td>
</tr>
<tr>
<td>Licensing Server (Citrix Cloud)</td>
<td>Manages licensing for Citrix products and services.</td>
</tr>
</tbody>
</table>
Citrix HDX 3D Pro for GPU based medical imaging applications

In this Reference Architecture, Citrix HDX 3D Pro is leveraged within the Citrix Virtual Apps and Desktops Service. Citrix HDX 3D Pro is a part of Citrix’s HDX remote display technology built on Independent Computing Architecture (ICA) which utilizes a graphics processing unit (GPU) for hardware acceleration. This solution uses NVIDIA cards; NVIDIA Encode (NVENC) technology which will offload the encoding of H.264 and H.265 video streams from the CPU and move this workload to silicon in the GPU resulting in substantially improved graphics for end users and higher overall user density.

Citrix Content Collaboration

Citrix Content Collaboration empowers mobile workstyle and software-defined workspaces by providing a secure file sharing & sync solution to the users. At the same time, it enables management and control to meet corporate data policies and unique compliance requirements. This solution perfectly fits into the healthcare model as it empowers physicians, nurses, and other medical professionals to collaborate effectively and efficiently by providing them a more straightforward yet secure way to access, share, and send files and data while embracing their mobile workstyle.

Citrix Content Collaboration has document workflows and a management module which can be leveraged in a healthcare environment to automate the file management for patients so that collaborative data and records can be accessed securely on the go by the medical staff.

The access is managed by the Citrix Files app, which allows secure data sharing and storage along with customizable usage and settings.

Citrix Content Collaboration architecture consists of two distinct regions – the Control plane and StorageZones. The Control plane is hosted by Citrix in the cloud and managed by Citrix as a service that is responsible for authentication, access control, brokering, and reporting.

StorageZones provide the data storage platforms for Citrix Content Collaboration, which can be customer-managed (on-premises or public cloud) or Citrix-managed (via the public cloud). Figure 22 depicts customer-managed StorageZones as this deployment utilizes HPE 3PAR’s performance, scalability, and data services for reliable and secure storage of critical healthcare data while maintaining sovereignty.

Figure 22. Authentication and data flow diagram of Citrix Content Collaboration
For detailed information about features offered by Content Collaboration visit https://www.sharefile.com/

Solution components

Hardware
Table 4 shows the hardware components used in this solution.

Table 4. List of Hardware components

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPE Synergy 12000 Frames</td>
<td>3</td>
<td>Infrastructure for compute, storage, fabric and management</td>
</tr>
<tr>
<td>HPE Synergy Composer 2</td>
<td>2</td>
<td>Composable Infrastructure management</td>
</tr>
<tr>
<td>HPE Synergy 480 Gen10 Compute Module</td>
<td>9</td>
<td>Compute Host</td>
</tr>
<tr>
<td>HPE 3PAR 8440 AF</td>
<td>1</td>
<td>Storage</td>
</tr>
<tr>
<td>HPE PCIe Expansion Module</td>
<td>4</td>
<td>Expansion Module to accommodate NVIDIA GPU</td>
</tr>
<tr>
<td>NVIDIA T4 GPU</td>
<td>16</td>
<td>GPUs dedicated for accelerating Citrix Virtual Apps and Desktops Service and Inference</td>
</tr>
</tbody>
</table>

Software
Table 5 shows the software components used in this solution.

Table 5. List of Software components

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPE OneView</td>
<td>4.2001.01</td>
<td>Infrastructure Management Software</td>
</tr>
<tr>
<td>Citrix Cloud</td>
<td>--</td>
<td>Cloud Platform for hosting Citrix services for Citrix Content Collaboration and Citrix Virtual Apps and Desktops Service</td>
</tr>
<tr>
<td>Citrix Virtual Apps and Desktops Service</td>
<td>1906</td>
<td>Citrix Cloud offering to deliver Citrix Virtual Apps and Desktops Service from on-premises and/or cloud Resource Locations</td>
</tr>
<tr>
<td>Citrix Content Collaboration</td>
<td>5.7</td>
<td>Secure file sharing &amp; sync solution which also allows collaboration</td>
</tr>
<tr>
<td>VMware vSphere®</td>
<td>6.7</td>
<td>Virtualization and Virtual Machine management platform</td>
</tr>
</tbody>
</table>

Best practices and configuration guidance

HPE Synergy configuration
Refer to HPE Synergy Configuration and Compatibility Guide to setup the hardware. The three (3) HPE Synergy 12000 Frames were configured with redundant HPE Synergy Composers, five (5) HPE Synergy 480 Gen10 servers have been used in the solution for non-accelerated EHR workload, four (4) HPE Synergy 480 Gen10 w/ PCIe Exp module have been used for accelerated Citrix Virtual Apps and Desktops Service.

Network
Configuration of VLANs is done via HPE OneView through HPE Synergy Composer, which is consumed by server profiles and logical interconnects.
Table 6 shows the list of VLANs considered in this solution.

<table>
<thead>
<tr>
<th>VLAN Type</th>
<th>VLAN ID</th>
<th>VLAN Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data center Management</td>
<td>100</td>
<td>Deployment</td>
</tr>
<tr>
<td>Infrastructure Management</td>
<td>22</td>
<td>Mgmt/Virtsw_net</td>
</tr>
<tr>
<td>Solution Network</td>
<td>223</td>
<td>VLAN_223</td>
</tr>
<tr>
<td>Fibre Channel 1</td>
<td>-</td>
<td>SAN_A_FC</td>
</tr>
<tr>
<td>Fibre Channel 2</td>
<td>-</td>
<td>SAN_B_FC</td>
</tr>
<tr>
<td>VMware HA</td>
<td>99</td>
<td>VMware HA</td>
</tr>
<tr>
<td>VMware vMotion</td>
<td>90</td>
<td>VMware vMotion</td>
</tr>
</tbody>
</table>

In addition to the above Ethernet Networks, a redundant Fibre Channel SAN network was also configured to connect vSphere with HPE 3PAR All Flash Storage Array as shown in Figure 23.

Figure 23. Network list in HPE OneView
Logical Interconnect groups

A logical interconnect group is added in HPE OneView for the Virtual Connect SE 40Gb Ethernet connectivity with the Interconnect Bay Set 3 & 6. Figure 24 shows the logical interconnect group created as 'Virtsw_LIG' where Virtual Connect SE 40Gb F8 modules are selected in a redundant manner. After selecting the appropriate interconnects, create an uplink set by selecting 'Add Uplink Set' and choose 'Type as Ethernet'.

**Figure 24.** Logical Interconnect Group
Enclosure Group
An enclosure group with enclosure count was created in HPE OneView. The ‘Virtsw_LIG’ is the logical interconnect group in the Interconnect Bay Configuration as shown in Figure 25.

Logical Enclosure
In HPE OneView, a logical enclosure is created that includes three enclosures along with the enclosure group. The logical enclosure applies the logical interconnect and logical enclosure configuration including the HPE Virtual Connect fabric.
Server Profile Template
A Server Profile Template needs to be created for two types of servers, i.e., HPE Synergy 480 Gen10 and HPE Synergy 480 Gen10 w/ PCIe Exp module. These templates are used to create Server Profiles which are applied to Servers. The template for HPE Synergy 480 Gen10 was applied to five servers and the template for HPE Synergy 480 Gen10 w/ PCIe Exp module was applied to three servers.

The network configuration for the Server Profile Templates is shown in Figure 27. Each server has two connections to the Solution Network VLAN and Fibre Channel SAN Network to ensure redundancy.

Storage
HPE Synergy was configured with a SAN Manager to manage the Fibre Channel Network connecting HPE 3PAR All-Flash Storage Array. Two SANs were created, each associated with a Fibre Channel connection. The Storage Pools in HPE OneView define different RAID types that are available on HPE 3PAR All-Flash Storage Array. Volumes were created as logical partitions and attached to the Server Profiles via the Fibre Channel network.

Figure 27. Connections in a Server Profile
Figure 28. Storage Volumes as in HPE OneView
On the HPE 3PAR StoreServ 8440 Storage, volumes were configured via StoreServ Management Console to Store User Profiles and Medical Imaging Data. The Volumes in HPE 3PAR StoreServ 8440 are created out of a Common Provisioning Group which represents a set of Disks configured with a RAID definition. In this solution, RAID 5 was used as it combines the better elements of efficiency and performance among the different RAID configurations.
VMware vSphere Server Configuration

Download the HPE Custom Image for VMware ESXi 6.7 from the HPE Support Drivers & Software or VMware Software site. A Hewlett Packard Enterprise custom image has all the necessary management tools for the Agentless Management Service (AMS) and Active Health Service (AHS). It enables drivers, utilities, and tools for device enablement.

VMware vSphere 6.7 should be installed on all HPE Synergy 480 Compute Nodes and associated with the required VLANs. For this test, local storage was utilized, but Hewlett Packard Enterprise offers numerous storage choices from HPE 3PAR StoreServ and HPE Nimble storage arrays to VMware vSAN. SSH is enabled on all the ESXi hosts to allow further configuration. To manage all the VMware vSphere ESXi hosts, either leverage an existing or create a new vCenter Server in the data center in which the infrastructure resides. All HPE Synergy nodes associated with the VDI solution should be housed in a cluster in a VMware data center within the vCenter server.

For this solution, two Clusters were defined. One cluster has five HPE Synergy 480 Gen10 servers for the non-accelerated workload and the other cluster has four HPE Synergy 480 Gen10 w/ PCIe Exp module for the accelerated workload. Both the Clusters have their own Distributed switch with a dedicated VMkernel adapter for vMotion®.

Refer to Deploying VMware vSphere on HPE Synergy guide for the deployment of VMware vSphere and refer to NVIDIA Grid deployment guide for Citrix XenDesktop on VMware vSphere for NVIDIA Grid cards configuration.

Preparing the Windows VM Template/Image

Perform the following steps to create a master template/image:

1. Create a Windows virtual machine and install the latest updates. House the VM in the Resource Cluster. This VM template will be consumed by Citrix MCS to create multiple Citrix Virtual Applications and Desktops Service.

2. Optimize the Windows virtual machine using the Citrix Optimizer. The tool should be downloaded on to the Windows VM to streamline the operating system by disabling unwanted features that can end up consuming more resources. The Citrix Optimizer provides the easiest and most efficient way to optimize your Windows desktop and server master images. It offers many advantages over traditional scripts, including the ability to roll back changes, selectively edit optimization values, and view detailed audit information before applying optimizations. The Citrix Optimizer includes customizable built in templates to enable or disable Windows features and system services across multiple platforms. Some of the best practices to be followed while using Citrix Optimizer include:
a. Use the Citrix Optimizer on a base image. Apply the tool to an unused system that has been built to match the configuration that you will deploy for virtual desktops or RDSH servers.

b. Disable as many unnecessary Windows OS components as possible. Consider disabling everything and then performing user-acceptance testing (UAT) to see if there are any issues. Re-enable any component that cause a problem when disabled.

3. The Windows 10 VMs used for accelerated workload testing were configured with NVIDIA profile “grid_t4-1b4”. Refer to NVIDIA Grid deployment guide for Citrix XenDesktop on VMware vSphere for the configuration details.

Citrix Virtual Apps and Desktops Service Configuration

Citrix Infrastructure was set up on Citrix Virtual Apps and Desktops Service delivery model with Citrix Cloud. The configuration required three Citrix Cloud Connectors, two Citrix StoreFront Servers, and one Citrix License Server. All these Servers were configured on Infrastructure Management VLAN. The Citrix Workspace 1906 was used on Endpoint VMs. The endpoint VMs were also placed in Management VLAN. The VDAs were Windows VMs, created using Machine Creation Service (MCS) on the Synergy Infrastructure in the Solution Network VLAN.

Citrix Cloud

Citrix Virtual Apps and Desktops Service was used in Citrix Cloud. This service hosts the Cloud Delivery Controllers used for authentication and brokering of Citrix Virtual Apps and Desktops Service. The local Domains were added under “Identity Access Management” in the Active Directory section in Citrix Cloud. The Access to the delivery groups was controlled via “Library” in the Citrix Cloud.

Citrix Cloud Connector

Citrix Cloud requires at least two Cloud Connectors and three Cloud Connectors is the best practice recommendation from Citrix. The VMs were sized according to the Citrix recommendations (4vCPU and 4GB RAM). A separate Resource Location was created for the three Cloud Connectors in the Cloud. The Cloud Connector Servers were Windows Server 2016 VMs and had Cloud Connector version 6.16.0.31167 installed.

Citrix StoreFront

The Citrix StoreFront was customer managed and version 1906 was installed on Windows Server 2016 VMs. Two Citrix Storefront Servers were used. The VMs had 8GB memory and 8 vCPU for each memory. A single Store was configured for launching both Citrix Virtual Apps and Desktops Service.

Citrix License Server

Citrix License Server Version 11.15.0.0 Build 27000 was installed on a Windows Server 2016 VM.

Machine Catalog, Delivery Group and VDA Configurations

The two (2) Machine Catalogs were created for this configuration. The VDA version used was 1906 and the catalogs were created without a Cache for Temporary Data. VMs in both the Catalogs was power managed by Delivery Controllers on Citrix Cloud.

The first Machine Catalog housed 35 Windows Server 2016 VDAs for the non-accelerated apps. Each VDA VM had 16 vCPU and 64GB RAM. Apps were delivered by a Delivery Group based on the applications that were published. The permissions to access the Delivery Group was managed from the Citrix Cloud Library.

Another Machine Catalog housed 183 VMs for the accelerated workload. These Windows 10 VMs had 4GB RAM and 2 vCPU each and were configured with NVIDIA drivers. Refer to NVIDIA Grid deployment guide for Citrix XenDesktop on VMware vSphere for the configuration details. These machines were delivered by a separate Delivery Group. The permissions to access the Delivery Group was managed from the Citrix Cloud library.

Capacity and sizing

As part of performance analysis and benchmarking, Hewlett Packard Enterprise set out to validate the Citrix Virtual Apps and Desktops Service powered by HPE Synergy Composable Infrastructure specifically for healthcare organizations. The goal here was to identify optimum user density which would result in maximum resource utilization without compromising on the user experience and system performance.

Login VSI

All performance testing documented utilized the Login VSI (http://www.loginvsi.com) benchmarking tool. Login VSI is the industry-standard load testing solution for centralized virtualized app and desktop environments. When used for benchmarking, the software measures the total
response time of several user specific operations being performed within a desktop workload in a scripted loop. The baseline is the measurement of the response time of specific operations performed in the desktop workload, which is measured in milliseconds (ms).

There are two values that are important to note:

- **VSIbase**: A score reflecting the response time of specific operations performed in the desktop workload when there is little or no stress on the system. A low baseline indicates a better user experience, resulting in applications responding faster in the environment.
- **VSImax**: The maximum number of desktop sessions attainable on the host before experiencing degradation in host and desktop performance.

Hewlett Packard Enterprise used Login VSI 4.1.39.6-Pro to perform the tests. The VMs were hosted on a single compute node for deriving the VSImax and VSIbase, maintaining a consistent number of VMs on each node. All virtual machines were powered on, registered, and idle prior to starting the actual test sessions.

Table 7 shows the summarization of the results from the tests.

<table>
<thead>
<tr>
<th>Component Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Accelerated Office worker workload</td>
<td>Testing applications related to Productivity, Images, Video, and Text</td>
</tr>
<tr>
<td>Accelerated worker workload</td>
<td>Testing GPU based applications related to 3D-Imaging, Scanning and image-based analysis</td>
</tr>
</tbody>
</table>

A standardized set of applications were installed on virtual machine for carrying out the performance benchmarking.

Table 8 shows the set of applications tested by Hewlett Packard Enterprise along with their versions.

<table>
<thead>
<tr>
<th>Application</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft Windows Server</td>
<td>2016 – 1607 (OS Build 14393.447)</td>
</tr>
<tr>
<td>Microsoft Windows 10</td>
<td>1607 (OS Build 14393)</td>
</tr>
<tr>
<td>Microsoft PowerPoint</td>
<td>2013 – 15.0.4569.1506</td>
</tr>
<tr>
<td>Microsoft Outlook</td>
<td>2013 – 15.0.4569.1506</td>
</tr>
<tr>
<td>Microsoft Word</td>
<td>2013 – 15.0.4569.1506</td>
</tr>
<tr>
<td>Microsoft Excel</td>
<td>2013 – 15.0.4569.1506</td>
</tr>
<tr>
<td>Google Chrome</td>
<td>78.0.3904.108</td>
</tr>
<tr>
<td>Doro PDF</td>
<td>1.82</td>
</tr>
<tr>
<td>Adobe Acrobat Reader</td>
<td>11.0.00</td>
</tr>
<tr>
<td>MicroDicom</td>
<td>3.01</td>
</tr>
<tr>
<td>3D Slicer</td>
<td>4.8.1</td>
</tr>
</tbody>
</table>
Figure 32. 3D chest scan image rendering using 3D Slicer

Figure 33 shows the samples of image rendering capability of the NVIDIA T4 GPU using the MicroDicom application. MicroDicom is used for primary processing and preservation of medical images in Dicom format.

Figure 33. Image rendering using MicroDicom
Non-Accelerated Office Worker workload
Hewlett Packard Enterprise validated the hosting of Citrix RDSH desktops on HPE Synergy 480 Gen10 nodes and utilized Login VSI for benchmarking.

Table 9 shows the non-accelerated office worker workload.

**Table 9: Non-accelerated office worker workload**

<table>
<thead>
<tr>
<th>Workload</th>
<th>vCPU</th>
<th>Memory</th>
<th>User</th>
<th>Apps</th>
<th>Video</th>
<th>Graphics Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Accelerated Office Worker</td>
<td>16</td>
<td>64GB</td>
<td>50/VM</td>
<td>Outlook, Word, PowerPoint, Excel, Chrome, PDF, 7-ZIP</td>
<td>480p</td>
<td>Not Present</td>
</tr>
</tbody>
</table>

Results for HPE Synergy 480 Gen10 server – Single Node
For a single HPE Synergy 480 Gen10 node an Office Worker VSImax of 296 user sessions was achieved. Seven Windows Server 2016 machines were functioning as Citrix VDAs on this Server. The graph shows the output of this test, the VSI Threshold represented by horizontal red line is 1558, which is crossed by Maximum Response (plotted red line) only near the end of the test when CPU resources are scarce. The Average Response (plotted yellow line) crosses the Threshold at the end of the test. This depicts that a single HPE Synergy 480 Gen10 node can easily support 296 users without degrading user experience.

![Figure 34. Office Worker on Single HPE Synergy 480 Gen 10 Node](image)
Results for HPE Synergy 480 Gen10 server – Five Nodes

For the combined run on five HPE Synergy 480 Gen10 nodes, the VSI Base was 519ms and VSImax was not achieved in any run. 35 Windows Server 2016 machines were functioning as Citrix VDAs on these five nodes. Each HPE Synergy 480 Gen10 node houses seven VMs. VSImax average was 673ms, VSImax threshold was 1520ms. Figure 35 shows the average response time, represented by VSImax average response (yellow line) which was consistently low. This indicates how well the HPE Synergy infrastructure handles the 1687 Office Worker users. The maximum response time represented in the chart by the plotted red line exceeds the VSImax threshold towards the end of the test. This is acceptable behavior and this situation is very much expected as the CPU resources in the cluster are exhausted. These values indicate excellent performance in each run as the point at which the VSI Index average (blue line) crosses the VSImax threshold (horizontal red line) is where the performance is deemed to be no longer acceptable by end users. The focus of the test is to ensure average response time stays within the VSImax threshold.

NOTE

The response time was within the VSImax threshold in every test that was performed.

Figure 35. Office Worker on HPE Synergy 480 Gen10 Nodes
Accelerated Office Worker workload with NVIDIA T4 GPU

Hewlett Packard Enterprise validated the hosting of Citrix Virtual Apps and Desktops Service on HPE Synergy 480 Gen10 nodes with four (4) NVIDIA T4 GPU and utilized Login VSI for benchmarking.

Table 10 shows Accelerated Office Worker workload with NVIDIA T4 GPU.

Table 10: Accelerated office Worker workload

<table>
<thead>
<tr>
<th>Workload</th>
<th>vCPU</th>
<th>Memory</th>
<th>User</th>
<th>Apps</th>
<th>Graphics Card</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accelerated Office Worker workload</td>
<td>2</td>
<td>4GB</td>
<td>1/VM</td>
<td>MicroDicom, 3D Slicer, Outlook, Word, PowerPoint, Excel, Chrome, PDF, 7-ZIP</td>
<td>1GB vGPU</td>
</tr>
</tbody>
</table>

Results for HPE Synergy 480 Gen10 server with NVIDIA T4 – Single Node

The total GPU memory available on each host is 61436 MiB Each Virtual Machine is allotted 1 GiB vGPU slice. The NVIDIA profile used for each VM is “grid_t4-1b4”. Therefore, each host can support up to 61 VMs with 1 GiB vGPU memory. Figure 36 depicts the CPU, Memory, and Disk utilization of the host when all 61 VMs are using the vGPU. The VSI Threshold, represented by horizontal red line is 1558 and the average response represented by plotted yellow line, never exceeds the VSI Threshold value, which is the indicative of superior performance for all the users.

Figure 36. Users with Accelerated Workload on HPE Synergy 480 Gen10 server with NVIDIA T4
Figure 37 outlines the GPU usage when 61 VMs are running with 1 GiB vGPU memory. The graph shows four plotted lines in Orange, Green, Blue, and Dark Grey color, representing the utilization of each T4 card. The maximum GPU utilization noted was 91% when all the VMs were running 3D workload (Application 3D Slicer). We see the graph tapering down in the middle as the 3D workload completes. Figure 37 shows the 3D Slicer and MicroDicom Workload used to test the GPU acceleration.

![Figure 37](image)

**Figure 37.** 3D Workload GPU utilization for NVIDIA profile “grid_t4-1b4” on 4x NVIDIA T4

Figure 38 outlines the GPU consumption when images with high resolution are opened with workload (Windows Photo Viewer). The graph tapers in the end, indicating GPU usage is going down after workload completion.

![Figure 38](image)

**Figure 38.** Image Processing GPU utilization for NVIDIA profile “grid_t4-1b4” on 4x NVIDIA T4
Results for HPE Synergy 480 Gen10 server with NVIDIA T4 – Three Nodes
For the combined run on three HPE Synergy 480 Gen10 nodes with four NVIDIA T4 GPU each, the total GPU memory available is 184308 MiB. These three hosts can support 183 VMs in total with each VM assigned 1 GiB vGPU. The below graph depicts the CPU, Memory, and Disk utilization for the combination of the three hosts when all 183 VMs are using the vGPU. The VSI Threshold (Horizontal red line) is 1522. The VSI max response (Plotted red line) does not cross the VSI Threshold at any point in time for these 183 users. The performance noticed was excellent for all the users.

![Graph showing CPU, Memory, and Disk utilization for three HPE Synergy 480 Gen10 servers with NVIDIA T4.]

**Figure 39. Users with Accelerated Workload on three HPE Synergy 480 Gen10 servers with NVIDIA T4**

**Note**
These graphs represent the results of tests run on the Citrix Virtual Apps and Desktops Service n-premises delivery model. These results can also be used to calculate approximate user capacity for the Citrix Virtual Apps and Desktops Service delivery model with Citrix Cloud. As defined in Software Components Figure 21 in the Citrix Virtual Apps and Desktops Service delivery model, the Control and Access planes are hosted on Citrix Cloud. The cloud-based Delivery Controller authenticates and brokers the sessions for users. Since these graphs define the performance of underlying hardware for the Resource Layer, they apply to both delivery models.

**Benchmarks versus field implementation**
Login VSI presents a relatively replicable set of tests that can be used to compare platforms and solutions within a fairly close range. The test uses a standardized set of workloads to create those comparison points. In the real world, it is highly unlikely that a customer will be running the exact set of applications featured in the test. As with most benchmarking tools, Login VSI results should be used in conjunction with results from actual system performance data from the field or via Proof-of-Concept (POC) implementations. The Login VSI presents response times from various tasks and applications that could be used as a primitive baseline in a controlled environment with limited applications and resource assignments. Although these metrics are useful when comparing systems with similar resource attributes, they can be misleading when used to extrapolate to real-world implementations. As a result, the numbers in this document are guidelines only. Historically, Hewlett Packard Enterprise has recommended sizing solutions at 60–65% of Login VSI numbers. This recommendation, however, is dependent on the fact that similar resource allocation is used as in the test results presented. Hewlett Packard Enterprise strongly recommends a complete analysis of the specific
user requirements prior to any VDI implementations, not implementations based solely on benchmark results. Customers new or inexperienced with VDI should undergo a deeper assessment of their environment, prior implementing VDI to make sure they attain the results they desire. If such an assessment interests you, contact your Hewlett Packard Enterprise account team for further information on our HPE Mobility and Workplace Services at hpe.com/us/en/services/consulting/mobility-workplace.html.

**Summary**

Hewlett Packard Enterprise and Citrix have collaborated to develop a joint solution to address the end-user computing infrastructure challenges faced by healthcare organizations, as modern healthcare workspace is undergoing a rapid digital transformation driven by both user’s requirements and modern 3D applications. This Reference Architecture guides healthcare organizations looking for client virtualization to enable a comprehensive, flexible, and efficient VDI solution accomplishing requirements of different workloads (accelerated and non-accelerated) and users.

This solution is based on the HPE Synergy Composable Infrastructure platform that helps in delivering promised benefits of client virtualization while overcoming many common challenges. The challenges are:

- Ability to start small and scale out in affordable increments—from pilot to production
- HPE Synergy 480 Gen10 Compute Module is a record holder for virtualized performance at 6, 8, 16, and 24 nodes, which translates to unmatched client virtualization performance for a superb end user experience
- Reduced downtime and maintenance window with true stateless computing
- High GPU density and freedom to choose multiple GPU types to accelerate the workload
- NVIDA T4 GPU is a versatile platform that delivers high-end graphics processing power for all types of 3D application, Citrix Virtual Apps and Desktops Service, and Inferencing
- HPE 3PAR StoreServ 8440 All Flash Storage delivers flexible I/O host connectivity without compromising performance, scalability, data services, or resiliency

The key takeaways include:

- A VDI implementation is an ideal approach to address compliance, regulations, and security requirements, which are imperative to any healthcare organization.
- Citrix Virtual Apps and Desktops Service improve clinical workflow by enabling clinicians to access confidential patient data on any device, over any network (wired, wireless, LTE). Apps and data securely follow healthcare staff as they roam
- With improved data security, data is stored in the data center rather than on endpoints. A centralized repository helps in Protected Health Information (PHI), clinical trial records, and patient financial data secure
- Ensure business continuity with highly available architecture to reduce downtime for business-critical applications
- Increase the availability of 3D imaging across healthcare organizations using centralized GPU-accelerated applications accessible on demand, which help in effective visualization by physicians

**Note**

The solution is capable of delivering pixel-perfect image quality but there are many factors around compliance that need to be considered for medical imaging use cases. This Reference Architecture is not an entire replacement to diagnostics workstations.
Appendix A: Benefits of HPE Synergy Composable Infrastructure in VDI/virtual environments

HPE Synergy Composer powered by HPE OneView is core to its composability story. HPE Synergy delivers orchestration via a single interface to discover, inventory, configure, provision, update, and diagnose the composable infrastructure in a heterogeneous environment. This Reference Architecture uses HPE Synergy Composer powered by HPE OneView, to demonstrate the on-demand composability feature of the HPE Synergy platform. This Reference Architecture aims to demonstrate the use of fluid resource pools to effortlessly compose and recompose a high-end graphics VDI environment and vice versa, on a single block of disaggregated compute, storage, and fabric infrastructure. It serves as a proof point for dynamically switching workloads from one use case to workloads in an entirely different use case within minutes by using HPE Synergy Composer. Figure 40 shows an HPE Synergy graphical composable flowchart depicting the scenarios and benefits on repurposing infrastructure based on the demand of resources in a real-time environment.

Figure 40. Flowchart showing on-demand composability using HPE OneView
Appendix B: Bill of materials

The following BOMs contain electronic license to use (E-LTU) parts. Electronic software license delivery is now available in most countries. Hewlett Packard Enterprise recommends purchasing electronic products over physical products (when available) for faster delivery and for the convenience of not tracking and managing confidential paper licenses. For more information, contact your reseller or a Hewlett Packard Enterprise representative.

Note

Part numbers are at time of publication/testing and subject to change. The bill of materials does not include complete support options or other rack and power requirements. If you have questions regarding ordering, consult your Hewlett Packard Enterprise reseller or Hewlett Packard Enterprise sales representative. Refer to hpe.com/us/en/services/consulting.html for more details.

Table 11 provides Bill of materials for reference purpose.

<table>
<thead>
<tr>
<th>Part number</th>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P9K10A</td>
<td>1</td>
<td>HPE 42U 600x1200mm Adv G2 Kit Shock Rack</td>
</tr>
<tr>
<td>871940-B21</td>
<td>9</td>
<td>HPE Synergy 480 Gen10 CTO Cmpt Mdl</td>
</tr>
<tr>
<td>804942-B21</td>
<td>3</td>
<td>HPE Synergy Frame Link Module</td>
</tr>
<tr>
<td>872957-B21</td>
<td>2</td>
<td>HPE Synergy Composer 2</td>
</tr>
<tr>
<td>794502-B23</td>
<td>2</td>
<td>HPE VC SE 40Gb F8 Module</td>
</tr>
<tr>
<td>K2P89B</td>
<td>12</td>
<td>HPE 3PAR 8000 1.92TB+SW SFF SSD</td>
</tr>
<tr>
<td>Q2513A</td>
<td>1</td>
<td>HPE 3PAR StoreServ RPS Service Processor</td>
</tr>
<tr>
<td>H6Y97B</td>
<td>1</td>
<td>HPE 3PAR 8440 2N+SW Storage Base</td>
</tr>
<tr>
<td>OK734A</td>
<td>4</td>
<td>HPE Premier Flex LC/LC OM4 2f 5m Cbl</td>
</tr>
<tr>
<td>P9Q43A</td>
<td>2</td>
<td>HPE G2 Basic 4kVA/(20) C13 NA/JP PDU</td>
</tr>
<tr>
<td>797740-B21</td>
<td>3</td>
<td>HPE Synergy12000 CTO Frame 1xFLM 10x Fan</td>
</tr>
<tr>
<td>P07351-L21</td>
<td>9</td>
<td>HPE Synergy 480/660 Gen10 Xeon-G 6254 FIO Kit</td>
</tr>
<tr>
<td>P07351-B21</td>
<td>9</td>
<td>HPE Synergy 480/660 Gen10 Xeon-G 6254 Kit</td>
</tr>
<tr>
<td>P00918-B21</td>
<td>108</td>
<td>HPE 8GB 1Rx8 PC4-2933Y-R Smart Kit</td>
</tr>
<tr>
<td>P00924-B21</td>
<td>72</td>
<td>HPE 32GB 2Rx4 PC4-2933Y-R Smart Kit</td>
</tr>
<tr>
<td>870753-B21</td>
<td>6</td>
<td>HPE 300GB SAS 15K SFF SC DS HDD</td>
</tr>
<tr>
<td>P01367-B21</td>
<td>9</td>
<td>HPE 96W Smart Storage Battery 260mm Cbl</td>
</tr>
<tr>
<td>804424-B21</td>
<td>9</td>
<td>HPE Smart Array P204i-c SR Gen10 Ctrlr</td>
</tr>
<tr>
<td>777430-B21</td>
<td>9</td>
<td>HPE Synergy 3820C 10/20Gb CAN</td>
</tr>
<tr>
<td>779218-B21</td>
<td>4</td>
<td>HPE Synergy 20Gb Interconnect Link Mod</td>
</tr>
<tr>
<td>838327-B21</td>
<td>2</td>
<td>HPE Synergy Y Dual 10GBASE-T QSFP 30m RJ45 XCVR</td>
</tr>
<tr>
<td>798096-B21</td>
<td>2</td>
<td>HPE Synergy 12000F 6x 2650W AC Ti FIO PS</td>
</tr>
<tr>
<td>804938-B21</td>
<td>3</td>
<td>HPE Synergy 12000 Frame Rack Rail Option</td>
</tr>
<tr>
<td>804943-B21</td>
<td>1</td>
<td>HPE Synergy 12000 Frame 4x Lift Handle</td>
</tr>
<tr>
<td>Part number</td>
<td>Quantity</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>-------------------------------------------------------</td>
</tr>
<tr>
<td>859693-B21</td>
<td>1</td>
<td>HPE Synergy Multi Frame Master1 FIO</td>
</tr>
<tr>
<td>804101-B21</td>
<td>8</td>
<td>HPE Synergy Interconnect Link 3m AOC</td>
</tr>
<tr>
<td>720199-B21</td>
<td>2</td>
<td>HPE BLc 40G QSFP+ QSFP+ 3m DAC Cable</td>
</tr>
<tr>
<td>861412-B21</td>
<td>6</td>
<td>HPE CAT6A 4ft Cbl</td>
</tr>
<tr>
<td>JH397A</td>
<td>1</td>
<td>HPE FF 5940 2-slot Switch</td>
</tr>
<tr>
<td>JH180A</td>
<td>1</td>
<td>HPE 5930 24p SFP+ and 2p QSFP+ Mod</td>
</tr>
<tr>
<td>JH183A</td>
<td>1</td>
<td>HPE 5930 8-port QSFP+ Module</td>
</tr>
<tr>
<td>ROW29A</td>
<td>12</td>
<td>HPE NVIDIA T4 16GB Computational Accelerator</td>
</tr>
<tr>
<td>P14255-B21</td>
<td>4</td>
<td>HPE Synergy 480 Gen10 PCIe x4 Expansion Module</td>
</tr>
</tbody>
</table>
Resources and additional links

HPE Reference Architectures, hpe.com/info/ra

HPE Composable Infrastructure, hpe.com/info/composable

HPE Composable Systems, hpe.com/info/synergy-ra

HPE Synergy, hpe.com/synergy

HPE Synergy Planning Tool, hpe.com/solutions/synergy-planning-tool

HPE 3PAR StoreServ storage, hpe.com/storage/3par

HPE Technology Consulting Services, hpe.com/us/en/services/consulting.html


To help us improve our documents, please provide feedback at hpe.com/contact/feedback.