Cisco Introduction to End to End Desktop Virtualization

Executive Summary

Target Audience

The target audience for this whitepaper is both IT decision makers who are responsible for evaluating desktop virtualization solutions, as well as IT support staff that need familiarization with virtualization concepts.

Introduction

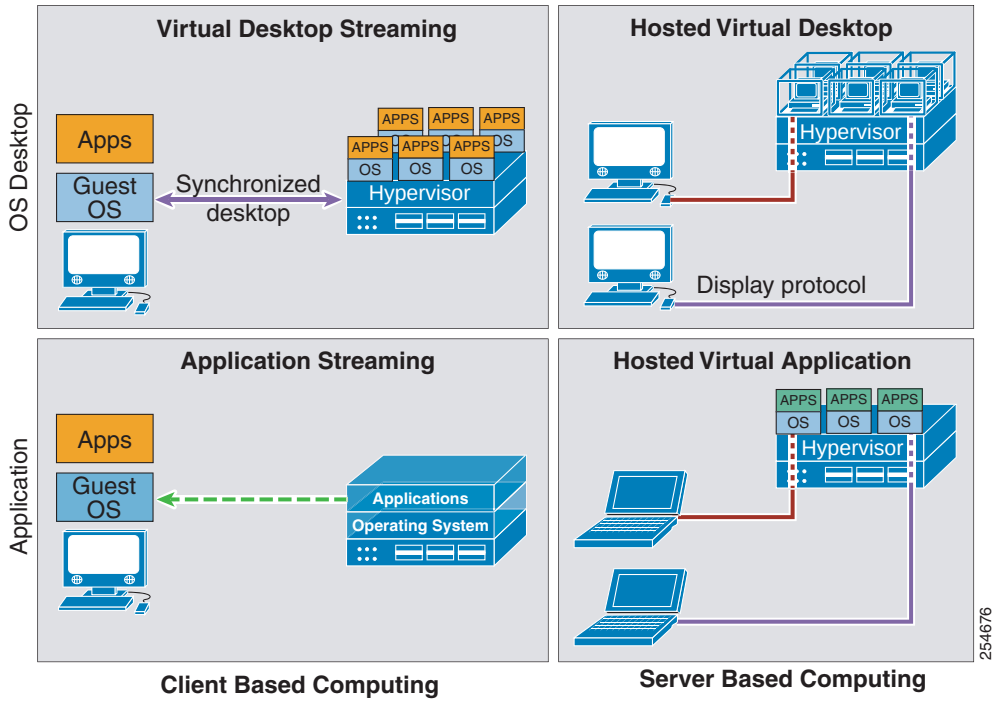
Today's PC-based enterprise desktop environment is highly distributed, giving users ample computing power and application flexibility. This desktop model works very well for the business needs of users. However, for IT departments, it presents multiple challenges such as high operating costs, complex management, and reduced data security.

Desktop virtualization (DV) technologies aim to solve the problems IT departments face while maintaining the current user experience. Multiple variants of DV are currently evolving in the marketplace, and each variant approaches DV from a different angle.

Four popular approaches are illustrated in Figure 1. These approaches differ according to whether processing takes place at the server or the client, and whether the complete desktop is virtualized or just the user's available applications. The choice of desktop virtualization drives client device hardware requirements, back-end software system complexity, network utilization, and user experience expectations.

Corporate Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Figure 1 Desktop Virtualization Technology Landscape



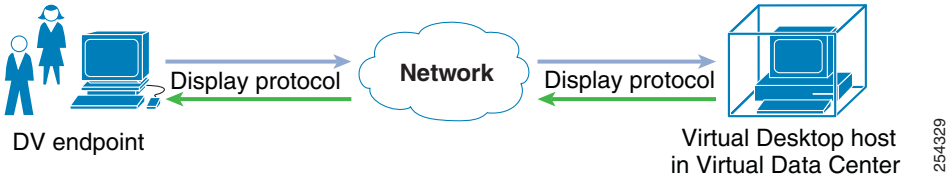
Hosted virtual desktops (HVD) and Hosted Shared Desktops (HSD) are growing in popularity since they meet most IT department requirements and maintain a local desktop-like user experience for the overwhelming majority of applications. All references to DV in this document refer to HVD DV unless otherwise stated. Please note that this chapter describes key DV-specific technology components and establishes the need for more. For a deeper understanding of each component and general installation discussions, please consult vendor-specific documentation.

Desktop Virtualization abstracts the end-user experience (operating systems, applications, and content) from the physical endpoint. With DV, the desktop environment is hosted in data center servers as virtual machines. Users can access their virtual desktops across the network by using laptop computers, thin client terminals, mobile Internet devices, and smartphones. Every large DV design should take into account user accessibility needs, specific user work profiles, applications, and the robustness of the underlying infrastructure. The infrastructure design should further consider the compute, storage, and network (data center, campus, WAN, or Internet) requirements. Each component plays an important role in delivering a good experience to the end user.

At a minimum, a DV session requires: an endpoint, a hosted desktop running on a virtual machine housed in a data center, and a software agent running inside the virtual desktop. The client initiates a connection to the agent on the virtual desktop, and views the desktop user interface with a display protocol. In a real-world scenario, another component called connection broker is present between the endpoint and the virtual desktop. The broker authenticates and connects the user to an appropriate virtual desktop.

As illustrated in Figure 2, there is a continuous exchange of display data between the virtual desktop in the data center and the endpoint, with a continuous exchange of interface device data (keyboard, mouse, USB peripherals, and so on) from the endpoint to the virtualized desktop. The network can be purely a LAN environment (for example, a corporate headquarters) or a mix of WAN and LAN environments. To allow for a consistent, high-quality, and scalable DV deployment, it is important to carefully plan the data center and the data network.

Figure 2 A Simple DV Session



User Experience with Virtual Desktops

Organizations considering desktop virtualization should begin by defining the primary attributes of good user experience. Depending on the type of user, actual DV feature sets may include a subset of the following features. The overall goal is to mirror the user experience in a DV environment so that it replicates the physical desktop experience as closely as possible.

User Input Response Time (Mouse and Keyboard Movement)

When a user moves the mouse pointer or types using the keyboard, the endpoint DV client sends these commands encapsulated inside the display protocol. Within the HVD, the DV software agent receives, decodes, and presents these commands to the guest operating system; the result of the commands, in terms of display changes, is sent back to the client over the display protocol. The compute capacity of the HVD, network latency, and bandwidth play a critical role in keeping the response time within a target value.

Application Response Time

Assuming user input response time is within the limits, application response time is governed by the amount of CPU, memory, and storage input/output per second (IOPS) seen by the virtualized guest OS. Application response time equivalent to what is seen by an end-user on a physical desktop is the target for all DV deployments. Setting very aggressive application response time expectations will significantly impact the density of HVDs in the data center. When comparing application response time on the physical desktop and on HVD, the OS versions and underlying hardware (physical or virtual) should be the same to achieve valid measurements.

Display Quality

The display quality seen by the end user is a function of the content viewed, screen resolution, and screen refresh rate. In general, for DV, high resolution and refresh rate result in higher compute and network loads. It is very important to assign enough computer resources and sufficient bandwidth based on user profiles and user location in the network. Most display protocols today support multiple monitors, either in span mode (single display sent from HVD and replicated to multiple monitors on the endpoint) or in true multiple-monitor mode (multiple displays sent from HVD, potentially each with a different screen resolution). Depending on the supported mode, the endpoint/HVD compute requirements will be impacted. In true multiple-monitor mode, network bandwidth requirements are much higher.

Interactive Multimedia

In enterprises today, interactive multimedia is critical for business productivity. Interactive multimedia incorporates everything from content-rich web pages with embedded video content, to collaboration and social network applications. The type of content may include Flash media, audio, video on demand, voice over IP (VoIP), application streaming, and so on. In a DV environment, interactive multimedia is typically requested by and rendered on the virtualized desktop in the data center. Once rendered on the virtual desktop, the media is re-encoded in the DV protocol and sent from the data center to the endpoint, where the DV protocol is decoded and finally displayed to the end user.

In this scenario, the interactive multimedia does not flow directly from the source to the endpoint; rather it is “hairpinned” through the data center and frequently tunneled through the DV protocol connection. Streaming media and real-time media are two primary means of interactive multimedia delivery, each presenting different challenges at the HVD in the data center.

As an example of streaming one-way media, when a user requests to view a web page containing flash video and audio, the following main events happen:

Display protocol consumes a small amount of network bandwidth to transport user request actions to HVD.

HVD consumes local compute, memory, and possibly storage resources to bring up a browser.

A connection to the web server hosting the web page is initiated (possibly over the Internet), consuming network bandwidth.

The content returned by the web server (such as images, flash video files and so on) is downloaded and processed by the browser, consuming CPU, memory, and storage resources.

The content is rendered on the HVD virtual display, captured by the DV display protocol agent, and transported to the endpoint. This consumes compute and memory resources, but more importantly network resources (LAN or WAN) are consumed a second time for the same content. DV display protocols are not as efficient at compressing video as codecs like MPEG4, so it is typical that the network bandwidth consumed to retransmit the video encoded in the DV protocol can be several times higher than the original compressed video stream.

Finally, the endpoint displays the video and uses the underlying audio drivers (if supported) to deliver the interactive multimedia.

Real-time media includes live video streams and two-way video/audio conference sessions. As an example, when a DV user requests to view a live video stream, the following events happen:

Display protocol consumes a small amount of network bandwidth to transport user request actions to HVD.

HVD consumes local compute, memory, and possibly storage resources to bring up a browser.

A connection to the content distribution server hosting the live stream is initiated (possibly over the Internet), thus consuming data center network bandwidth.

The live video returned by the CDS can’t be cached and is decoded using appropriate video codecs in the user’s HVD, consuming CPU and memory resources.

The content is rendered on the HVD virtual display, captured by the DV display protocol agent, and transported to the endpoint. This not only consumes compute and memory resources, but also, and more importantly, network resources (LAN or WAN) a second time for the same content. DV display protocols are not as efficient at compressing video as codecs like MPEG4, so it is typical that the network bandwidth consumed to retransmit the video encoded in the DV protocol can be several times higher than the original video compressed stream.

Finally, the endpoint displays the video and uses the underlying audio drivers (if supported) to deliver the interactive multimedia.

In a two-way video conference session, end-to-end audio latency and video rendering requirements are very stringent. In a user's HVD, the process of receiving the content, processing it, and preparing it for transport over the display protocol will add extra latency. For acceptable user experience, end-to-end latency should not exceed a few milliseconds. The HVD will need to have a sufficient CPU, memory, and network resources. To achieve this, sync between audio and video streams needs to be very tight. For this reason, extra compute resources in the HVD and strict QoS guarantees in the network are required.

Depending on the quality and complexity of the video and fidelity of the audio, the resources utilized can vary significantly. For most interactive multimedia content types, the general process for fetching and delivering the content to the end user essentially remains the same. The resources used for interactive multimedia delivery in a DV environment can be many times greater than in a traditional desktop environment, and this resource utilization is spread across multiple components. Further, scaling the model just described for thousands of geographically dispersed user groups requires careful capacity planning as well as a focus on optimization. Advanced network features such as multicast for streaming media, or video on demand, can't be applied in this model, since the media is contained in the display protocol and the display protocol must be point-to-point. Currently, the optimization techniques either use compression of the display protocol to conserve network bandwidth or separate the interactive multimedia from the display protocol itself (this is called multimedia redirection or MMR). MMR enables fetching and rendering of media at local endpoint and provides the opportunity to take advantage of advanced network features such as multicast, caching, targeted compression, and so on.

The technology to support a better interactive multimedia experience over high latency/low bandwidth networks and WAN is still evolving, but with careful planning, most day-to-day business needs can be met with the technology currently available.

Desktop Authentication

A user transitioning to a virtual desktop would expect to log in once on the endpoint and be served with the desktop. Authentication in a DV environment is required on the endpoint, on the connection broker, and finally on the HVD to get to the desktop. Since the connection broker integrates with existing Active Directory infrastructure securely in the data center, desktop authentication is easy to provision with negligible network or compute impact. User authentication information is passed to the connection broker from the endpoint over a secure connection inside the display protocol.

Note Applications within the HVD are separately authenticated and those mechanisms are out of scope for this discussion.

Local Devices

Local device and peripheral support in a DV environment play an important role in preserving the user's existing desktop experience. Some peripherals are an integral part of the day-to-day business and user experience, such as USB- controlled devices, printers, and access to local drives. Integrating these peripherals into the DV system can significantly impact network and compute requirements. The following need to be considered:

- Which users receive access to which peripherals?

Access is typically controlled by Active Directory policies, connection broker policies, and choice of endpoint hardware. Active Directory group policy-based control applies to user permissions inside the HVD itself. The endpoint is controlled by the software client, which in turn talks to the connection broker to procure specific policies. These policies can control access to USB devices, ability to copy data in and out, or ability to bypass display protocol. Most connection brokers available have the ability to derive these policies from Active Directory group policies.

- What is the location of these user groups?

  Location helps determine which network segments are impacted. Most local peripherals were designed with no network limitations in mind, and generally have ample hardware bandwidth (through a Peripheral Component Interconnect Express [PCIe] expansion card) and local compute resources for operations. When these local peripherals are made available to a HVD in the data center, limited network bandwidth and higher latencies may be introduced. For example, providing USB and local drive access to campus users on the LAN, has less impact on the network when compared to providing such access to WAN users.

Depending on the data collected, certain display protocols might work better than others and should be considered in the choice of any peripheral optimization software.

USB Redirection

Many new peripherals, like USB attached printers, webcams, audio headsets, or scanners, use the USB interface on the local device to connect. For these to work in a DV environment, the DV client on the endpoint intercepts the USB data and sends it within the display protocol to the agent installed on the HVD. The agent presents the USB device and data to the HVD as a new hardware device. Raw USB data streaming, depending on the software and hardware version of local USB controller, can consume 480 Mbps of network bandwidth and CPU cycles on the endpoint as well as in the HVD. Such high-bandwidth requirements can't work over a WAN link without optimization and compression on the local client itself. Available display protocol software clients (like VMware or Citrix) can compress and optimize the USB data to be sent over the network. Note that not all operating systems support compatible drivers, and a thorough compatibility check between the HVD OS, local client OS, DV client/agent software, and the end application is highly recommended. When using USB redirection over a WAN link, make sure enough bandwidth is provisioned and optimization is employed.

Printing

Printing is an important aspect of user experience, and it also has a significant impact on the performance and stability of any Windows-based HVD environment. In Windows-based printing, the application initiating the print process creates a printer-independent metafile, called an enhanced meta file (EMF), and sends it to the Windows spooler. The spooler then converts the EMF into a printer-specific raw file based on specific printer drivers and sends it to the USB or network-attached printer. In DV environments, all of these actions take place in the HVD, while the printer is present locally near the endpoint. As an example, a typical EMF file of 2 MB can expand to a 10-MB raw file. Given the amount of print traffic per user, it is highly recommended to optimize print traffic in a DV environment. Desktop virtualization solutions may use caching and compression to reduce raw print traffic.

Local Drives

Some users may need access to local drives for their jobs. Most connection brokers and display protocols support local drive mapping from the endpoint to the associated HVD. Active Directory group policies can also be used to control the usage. However, it is recommended that this be done on a case-by-case basis since one of the primary drivers for DV deployment is data security. The network challenges for a WAN environment remain the same as for any other local device mapping.

Display Protocols and Connection Brokers

The primary functions of display protocols are to transport display and user input data to and from the HVD and an endpoint. Other functions such as video optimization, audio transport, and USB/print data transport, are also performed by display protocols. The client (on the endpoint) and the agent (in the HVD) software-initiate and terminate the connection, respectively, and are responsible for interfacing with the OS drivers to produce a satisfactory user experience.

CPU and memory utilization on both the client and HVD are impacted by the choice of display protocol. Display protocols generally consist of multiple channels, each carrying a specific type of data stream. Advanced algorithms to optimize compute resource consumption and network bandwidth are built in many of the newer display protocols, clients, and agents.

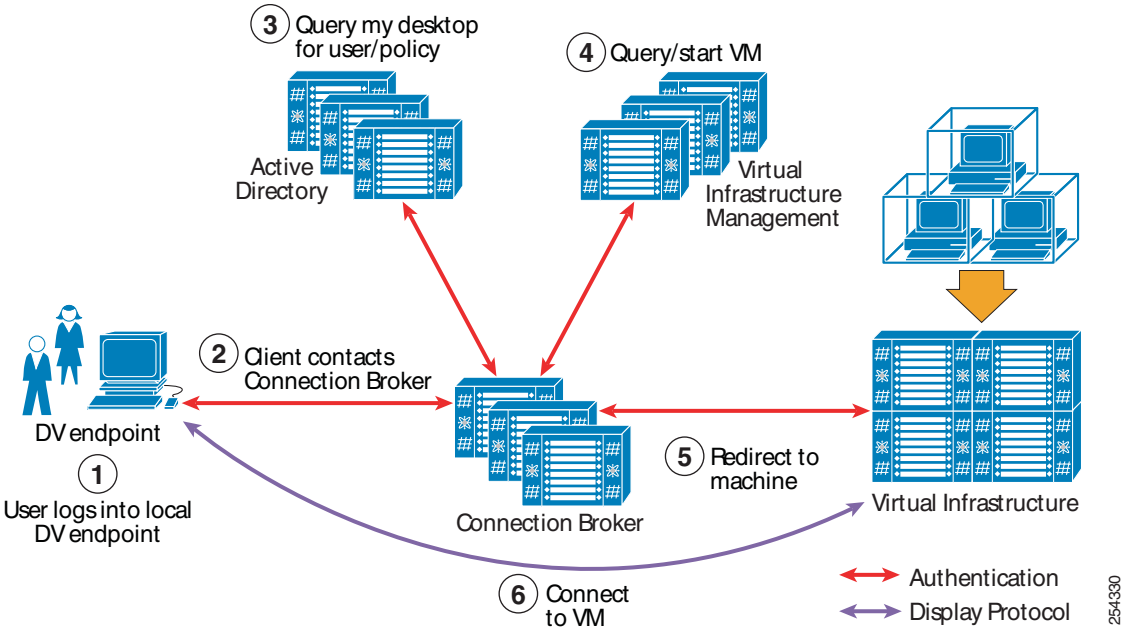
Display protocols currently present in the market are differentiated by how they deal with interactive multimedia. Variations include support from multimedia redirection to advanced progressive display rendering technologies. Three of the most commonly used display protocols are Independent Computing Architecture (ICA), PC over IP (PCoIP), and Remote Desktop Protocol (RDP). Some endpoint types, like Zero clients, support only a subset of display protocols. The choice of display protocol will heavily influence the way the network is configured, the amount of computation power required, policy application, future compatibility and most importantly, the user experience.

A lot of the added features already described are a function of the display protocol and the capabilities of the connection broker. A direct connection from an endpoint to the virtual desktop as depicted in Figure 2 is not scalable for a large deployment. As a primary DV component, the connection broker is required to authenticate and redirect multiple client connection requests to the virtual desktops. Typically, the connection broker also provisions new virtual desktops on demand, and relays user credentials to the hosted virtual desktop.

Advanced broker features allow multiple desktops per user, tunneled and proxy display protocol modes, support for creating large clusters with load balancing capabilities, and termination of SSL connections, to name a few. The broker typically communicates with the existing Active Directory infrastructure to authenticate users and to pull and apply user-specific desktop policies. The connection broker also maintains the state of the connection in case of drops or disconnects, and can optionally power down or delete the remote desktop.

Figure 3 illustrates connection flow in a simplified DV environment that includes a single endpoint and connection broker. This basic connection flow is present in all DV deployments, across all vendors. Certain vendors support tunneling modes on their connection brokers that vary in capabilities. Tunneling configurations have not been validated in this design guide and are generally not recommended in most scenarios. Tunneling requires termination and reorigination of display protocol sessions between user endpoint and HVD, causing an increase in the compute and memory required to support a single connection on the connection broker. In some deployment scenarios, for example, where Remote Desktop Protocol (RDP) over HTTPS is required for users accessing their HVDs over the Internet, dedicated tunneled connections might be needed and should be factored into the design.

Figure 3 Connection Flow



The following steps list the DV connection flow depicted in Figure 3 above:

Step 1 User logs onto the endpoint and accesses the client software.

Step 2 The client software initiates the connection to the broker.

Step 3 The broker looks up the user in the Active Directory and determines the associated remote desktop.

Step 4 The broker may optionally contact the virtual infrastructure manager to create/start the user's remote desktop.

Step 5 The user is then redirected to that remote desktop.

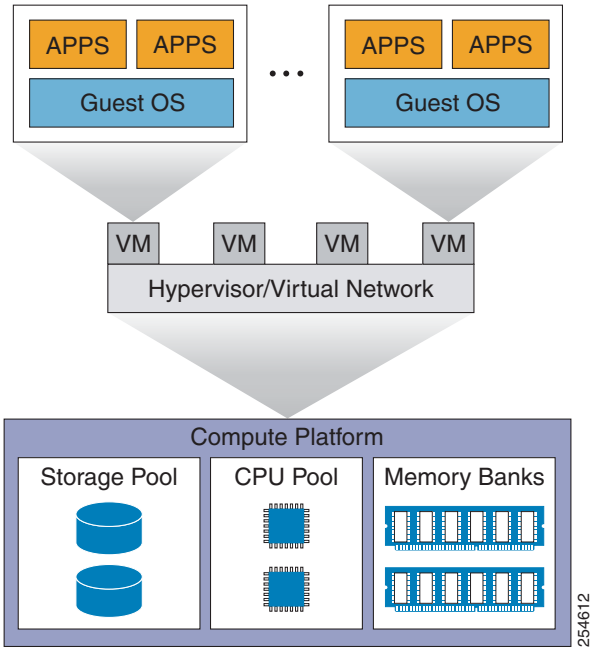
Step 6 The endpoint is thereafter directly connected to remote desktop via the display protocol.

Hypervisor

Virtual machines are isolated and abstracted virtual hardware containers where a guest operating system runs. The isolated design allows multiple virtual machines to run securely and access the hardware to provide uninterrupted performance. The layer of software that provides isolation is called the hypervisor. Figure 4 shows the hypervisor's role in the HVD.

Apart from hardware abstraction, a properly configured hypervisor ensures that all VMs receive a fair share of the CPU, memory, and I/O resources. The hypervisor has minimal or no direct control of the guest operating system running inside the VM itself. It does, however, indirectly control the OS behavior by controlling memory utilization, sharing CPU cycles, and so on. A virtual desktop operating system like Windows 7 runs inside a virtual machine, which in turn runs on the hypervisor-controlled host server. Cisco VXI support VMware ESX/ESXi, Citrix XenServer, and Microsoft Hyper-V hypervisors. Hypervisors typically provide a wide range of tools to achieve optimum virtual desktop densities in the data center.

Figure 4 Components of a Hosted Virtual Desktop



All virtual machines share common hardware resources on a single host (server) or in a pool of hosts (servers). In a typical DV deployment with a high density of virtual desktops running on a single server, it is highly advised that resource sharing features are used appropriately in the environment.

Resource sharing includes fair or configured distribution of hardware resources on a single host, identification of overcommitted host machines, live movement of VMs among hosts in a pool, VM power-up sequences based on priorities, and so on. In a typical DV scenario, the provisioned desktops are generally not all powered up at the same time. This is fundamentally different from the usage pattern of the typical always-on mode of operation of virtualized servers in the data center. Furthermore, virtual desktops in an enterprise environment are normally not used after hours and during holidays. Using the hypervisor's resource sharing capabilities along with power management features will reduce the data center's power usage footprint, especially in a DV deployment. The hypervisor's power management tools can consolidate all the active virtual desktops on the fewest possible hosts (servers), or power-down idle hosts, power-up hosts based on configured priority, and so on.

Both the resource sharing and power management features need VM migration capabilities, preferably live migration, and shared VM storage. Live migration enables movement of running virtual machines from one physical server to another with zero downtime, continuous service availability, and complete transaction integrity. The ability to move desktops within the data center based on predefined power and resource usage policies is extremely useful. Figure 5 illustrates the concept of resource sharing with VM Migration.

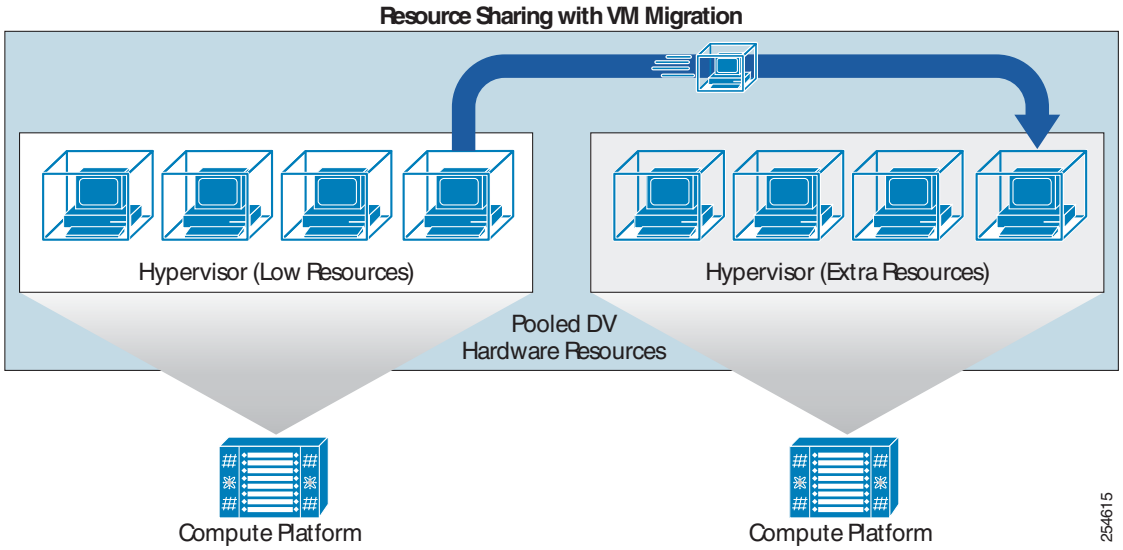
Hypervisors may implement a software-based virtual switch to which each HVD connects. The virtual switch presents virtual network interface cards (NICs) to the HVDs and connects to the physical NICs on the host machine. All data from the HVDs is forwarded by the virtual switch with appropriate VLAN tags and QoS policy application (if available).

Virtual switches implement basic switching features that in many scenarios are not sufficient in a real deployment. Citrix XenServer, for example, allows creating multiple networks, VLANs, physical NIC bonds with load balancing, and dedicated storage NICs. The VMware vSphere vSwitch provides all these same features, and with vCenter management service also provides a distributed virtual switch (DVS) functionality. The DVS functionality allows multiple VMs across multiple hosts to be connected

to a shared virtual switch. This feature makes it very easy to manage network configurations for large quantities of HVDs in a DV environment. Citrix XenServer uses an external switch platform to provide functionality similar to a distributed switch. A distributed switch enables the movement of VMs between hosts, which is a basic requirement for advanced features such as resource scheduling, power management, VM migration and HA discussed above. Although distributed virtual switches provide basic functionality, they do not provide comprehensive feature depth. For example, VMware's DVS feature lacks advanced networking features such as advanced QoS, stacked VLANs, 802.1x and port security that any access switch should possess. The Cisco Nexus® 1000V Series Switches provide all this and more.

In addition to the feature sets, a fundamental limitation of a virtual switch is that it consumes CPU cycles on the host machines themselves. If there is a lot of inter-VM communication, the host CPU consumption will be higher. In a high-density DV deployment, this resource utilization must be included during capacity planning. Further, if virtual switches packaged in the hypervisors are used, visibility into all inter-VM communication on a single host is lost. Certain types of security breaches, such as DoS attacks or even broadcast storms caused by misconfiguration, can have severe implications on the DV environment and are difficult to troubleshoot without visibility into the flows.

Figure 5 Resource Sharing with VM Migration



Note Currently, Cisco Nexus 1000V Series is only supported on the VMware ESX/ESXi hypervisor.

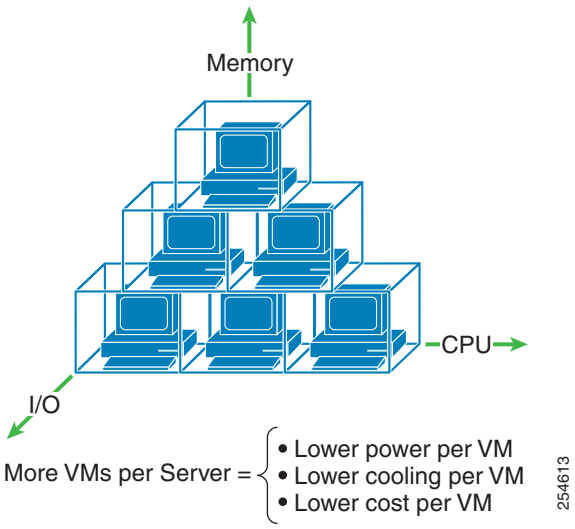
Hosting the Virtual Desktops

Whether DV is deployed in existing data centers or is greenfield, some important design criteria need to be considered. An enterprise DV environment requires scalable, secure, and always-on compute, storage, and network infrastructure. Specific guidelines depend on the DV vendor of choice, but the following general criteria specific to any DV deployment need to be considered.

As depicted in Figure 6, compute resources, memory and I/O form a decision axis for DV in the data center. Running out of one resource before the other for example, CPU, would waste other provisioned resources, memory, and I/O. At the same time, if the capacity planning is not appropriately done, it can

be difficult to meet the goals of obtaining the maximum density of HVD while preserving user experience and avoiding wasted resources. The next sections take a look at some of the fundamental differences in regard to the consumption of these resources in a DV environment.

Figure 6 Performance Dimensions



Compute

Server CPU power is many times that of a local desktop. That means a single multicore CPU can process multiple HVDs. Since applications in a desktop are specific to the user, the amount of memory sufficient for each HVD can be difficult to estimate. A DV deployment is more memory-centric than a virtualized server deployment. With most common workloads, maximum HVD density can be achieved by choosing a compute platform that can provide large memory banks.

Network

For display traffic, many elements can affect network bandwidth, such as protocol used, monitor resolution and configuration, and the amount of multimedia content in the workload. Concurrent launches of streamed applications can also cause network spikes. For WANs, you must consider bandwidth constraints, latency issues, and packet loss. In a DV environment, the density of virtual desktops supported by a single compute and network infrastructure is much higher than the traditional campus or branch infrastructure. The aggregation of endpoints increases network and compute load especially during peak usage periods.

The following primary data flows exist in a DV environment:

1. Display protocol traffic between HVD in data center and endpoint.
2. Traffic between HVD in the data center and the Internet. This traffic is generally routed over the enterprise campus network.
3. Storage data traffic (I/O per second [IOPS]) between HVDs and storage arrays. This traffic normally stays within the data center.
4. Application data traffic. This traffic generally remains within the data center.

All the above (with the exception of number 4) are new traffic flows introduced because of DV. In a large-scale deployment, each data flow can congest entry and exit points in the data center network. Appropriate QoS mechanisms to protect against delay and dropping sensitive data flow, such as display protocol data, is required.

Additionally, user usage patterns are unpredictable and may negatively impact bandwidth consumed per virtual desktop. For example, a high-quality, full-screen video can suddenly increase bandwidth consumption for a particular HVD multifold data center. Network capacity may normally be sufficient, but a careful design of the complete end-to-end network is required to preserve the user experience, and to allow for future expansion and accommodation of technology advances in DV space. In traditional enterprise desktop deployments, user desktops have one point of connection to the access network. This access layer separates the functions of the network and the actual compute resources cleanly, such that the policy application for network and user work environment do not conflict. In a DV environment, the endpoint still attaches to the campus access layer, but now there is also a HVD in the data center server attaching to the network at a virtual switch port inside the server. The addition of a new attachment point for the single user, and location of the virtual switch, blurs the boundary for policy application and opens new security concerns. Such concerns can be handled by either moving the attachment point of HVD to a controlled access switch, or by using a feature-rich virtual switch to shift the control inside the server itself.

Storage

A DV endpoint may have no local storage, and the associated HVDs are located on a centrally managed storage system. Normally, storage for each HVD is split into desktop stores and user profile stores. The split in desktop and user profile storage enables deletion of the desktop while still preserving the user profile; this, in turn, supports persistent and non-persistent desktops, data backup (user data backup only), and reduction in storage requirements (desktops are highly redundant and cloning can be used). It should be noted that while using non-persistent desktops reduces the storage capacity required for the HVD images, it does not change the amount of IO throughput required. This leads to smaller savings in terms of the number of storage disks and controllers needed to achieve the high IO throughput. A solution to reduce the redundant IO generated by the hypervisor and increase storage efficiency by employing intelligent caching solutions is highly recommended. Average storage I/O generation in a DV environment is predictable over a period of time, but very unpredictable in short duration. This is quite unlike server environments. However, IOPS is very dependent on the type of OS, its version, and kind of workload applications. Duplicate data in a DV environment is very high (common OS code base, similar applications installed), so DV storage can benefit enormously from data-de-duplication technologies and VM cloning. Storage requirements for each desktop can keep growing unless appropriate control policies are in place. In the case of cloned HVDs, it's highly desirable to restart the HVD periodically (to purge accumulated temporary data) and use disk quotas for users or assign controlled network storage instead of direct disk access. DV deployments are generally storage technology agnostic, and so storage area networks (SANs) or network attached storage (NAS) can be used, based on organizational preference.

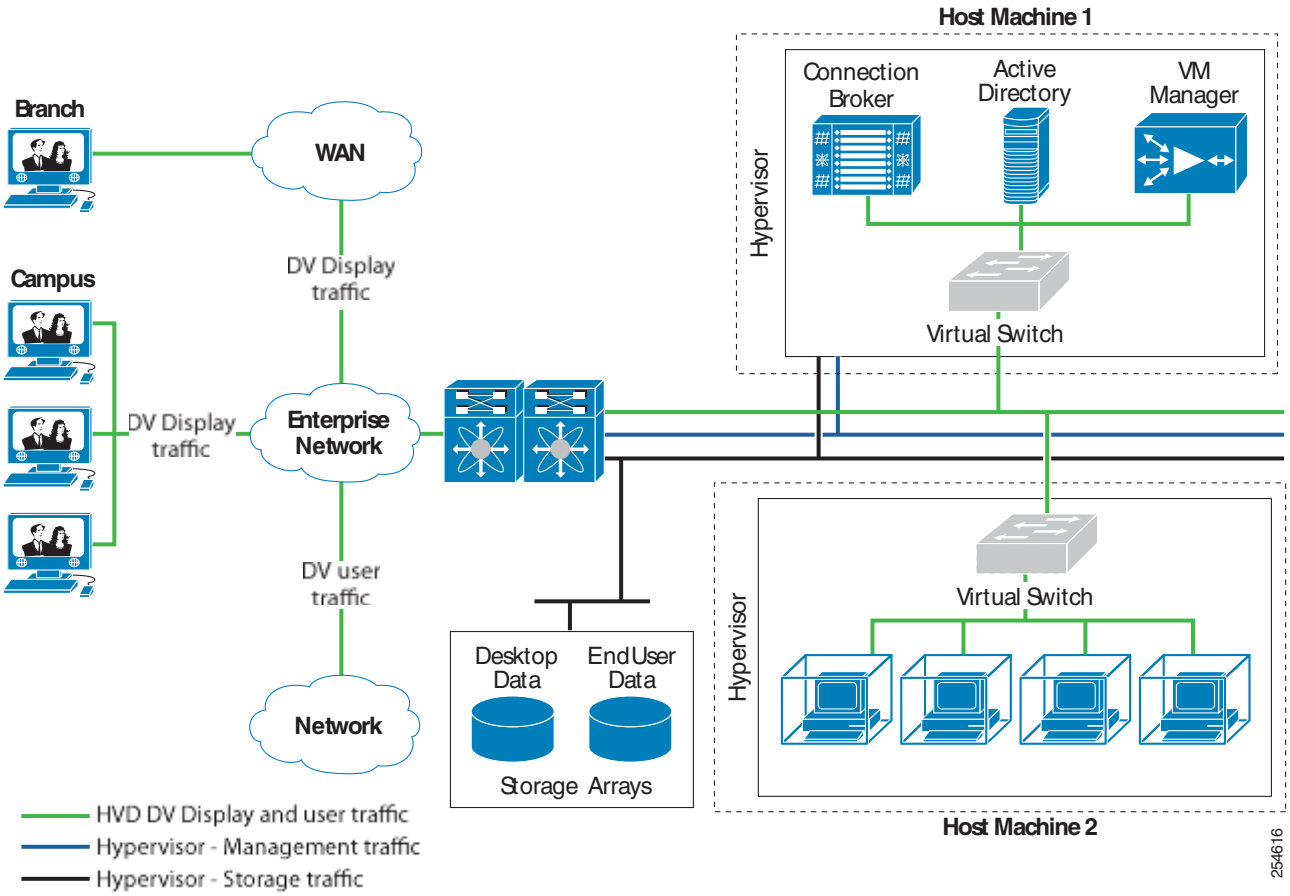
Typical Desktop Virtualization Deployment

Figure 7 depicts a typical DV deployment with all major components and connection paths. For DV, the data center houses, at a minimum, an HVD pool, connection broker, Active Directory, and virtual manager servers. Each HVD in the pool is a VM, and the pool itself is a physically separate host machine. Another host houses the VMs on the connection broker, Active Directory, and the VM manager that are installed. It is highly recommended to physically house HVD pools and enterprise services on different host machines. The hypervisors in each host connect to separate storage and management

networks marked black and blue, respectively. Note that the endpoints in the campus or the branch connect to the already existing access network infrastructure, and the HVDs in the pool attach to the virtual switch installed inside the hypervisor. All the endpoints, HVDs, and DV server components are on the "green" network. Devices in the green network can reach each other using Layer 2 or Layer 3 connectivity. All the display protocol traffic originates and terminates between the endpoints and HVDs.

Hosted Virtual Desktops can be deployed in multiple modes, such as dedicated, pooled, streamed and shared. These modes are differentiated based on how end users access and associate with the virtual desktop. Another way to differentiate HVDs is based on the state of virtual desktops across multiple user logons where the two modes are persistent one-to-one or non-persistent desktops. Persistent desktops are tied to a specific user, generally have customizations and user data preserved through the life of the desktop. Non-Persistent desktop pools allow better scale by creating an oversubscribed pool of desktops shared across users. These desktops are assigned to the user at logon, customizations and user data, if present in a separate data store, are applied and when the user logs off, the desktops are released in the pool for other users. Non-Persistent desktops are best suited for task workers who require temporary compute platform to access various enterprise applications, and are not expected to store user data locally. Main components for Hosted Virtual desktops and their deployment are discussed next.

Figure 7 DV Connection Paths



Note Figure 7 shows basic DV connectivity. For specific scenarios and features, the deployment could differ.

Note Figure 7 places the use of critical Cisco VXI components in the perspective of the network and does not illustrate all Cisco VXI components required to enhance user experience.

Workloads, Peak Workloads, and Design Impacts

Workload

For each user profile defined, granular workload parameters need to be set based on experience with existing physical desktops, or by testing. Workload parameters, like user profiles, are very unique to each enterprise. Early determination of these parameters makes the difference between the success or failure of a DV deployment. Workload parameters should include application run time, application usage patterns, average screen refresh rates and OS and application memory, CPU and storage footprint. Events that generate CPU, memory, or network load in the HVD, such as saving a document, sending an email, viewing a video, or editing text, need to be listed in a workload definition along with estimated average utilization. Each type of workload can have varying storage throughput requirements measured in IOPS, and care should be taken to discover these requirements very accurately, especially if planning a large scale DV deployment. Much of this data is generated during the pilot DV deployment phase and is then used to estimate resource requirements for full-scale deployments. Minor inaccuracies in estimating network, IOPs, CPU, and memory loads during pilot deployments can add up significantly in the final deployment.

Peak Workload

Because HVD runs on shared compute resources abstracted by the hypervisor, the resource utilization of one HVD can impact the others if appropriate hypervisor mechanisms and planning are not used during deployment. Since all users access their HVDs centrally located in data centers, they share a common secure network entry point into the data center. If not designed properly, this common network entry point can be a potential bottleneck, and can cause user experience degradation or even complete loss of connectivity. The same peak workloads can cause a storage bottleneck that degrades desktop performance. Identifying potential peak workload scenarios and sizing the network and storage system to handle peak workloads during DV deployment planning is essential.

As an example, consider an enterprise where hundreds of users start their HVDs in the morning, start Microsoft Outlook, and launch their productivity applications in a span of a few minutes. This sequence of events can cause severe strain on every component in the DV environment and possibly non-DV applications sharing the network. Since all user traffic is encapsulated in a display protocol, synchronized user actions, such as hundreds of users simultaneously watching a live corporate video, can congest the network. If mission-critical applications are consumed concurrently by a group of these users, there are no QoS mechanisms built within the display protocol to treat video and critical application traffic separately. Simultaneous antivirus checks, application and OS updates or HVD reboots, are other examples of peak workload.

Proper capacity planning, sufficient spare resources, intelligent use of resource scheduling, fault tolerance and high-availability (HA) features, and network capacity and QoS planning are absolutely essential to deal with peak workloads in a DV environment. Further, optimum WAN/Internet bandwidth utilization should be carefully considered and planned.

Powered with a clear understanding of average workloads, peak workloads, and user profiles, it's possible to make a good assessment of resources and configurations required in a particular DV deployment. Please see the discussion of capacity and performance in this document for more on making resource assessments with sample workloads.

Enterprise Policy and User Profiles

User profile definition is one of the first and most important steps when planning a DV deployment. User profile definitions are driven by an overall enterprise IT policy and business drivers behind DV deployment. Today's enterprise desktop virtualization decisions are primarily driven by data security, IT environment flexibility and total cost of ownership. The important policy decisions that define the DV deployment are listed next. Please note that this list is not comprehensive and specific business needs might require further policy granularity.

- Deciding user profiles and the total number of users to be migrated to DV:

  User profiles are generally categorized into task workers, knowledge workers, and power users. Task workers perform specific functions, with minimal requirement to change their desktop environment (for example, storing files locally, desktop personalization, and so on). Task workers normally do not require persistent desktops (enabling desktop sharing) and are ideal candidates for large scale DV. Knowledge workers are users who create content, consume rich media and in some cases need to install applications on their desktops (examples include product managers, executives, and software programmers). They normally require persistent dedicated desktops that allow the user's virtual desktop state to be maintained across logons . Power users are workers who need access to very high computing power for their day-to-day work. Power users manipulate large amounts of data, for example, graphic designers, CAD designers, architects, and so on. For a successful DV deployment, it is critical to crisply define user profiles based on the enterprise policy and business needs.

- Identifying business-critical applications required:

  Application requirements heavily drive storage, compute, and network requirements in a DV environment. These requirements can also help decide specific virtualization vendors and virtualization software to be deployed in the environment. Depending on the business requirements, applications could include SAP business software, customer relationship management (CRM) applications, software as a service (SaaS), web browsers, Microsoft Office, Microsoft Outlook, and so on. Generally speaking, any existing IT desktop policy should apply to applications deployed in the DV environment.

- Grouping of users based on geographical locations and places in the network:

  The location information defines which data centers the DV users are served from and also helps ascertain network, security, and localization policy requirements.

- Defining level of access per user or per user location:

  A large enterprise with diverse workforce (for example, fixed users, mobile users, telecommuters, contractors, and so on) and an existing IT infrastructure has well-defined data access policies in place. Most of these policies can be transplanted seamlessly into the DV environment. Keep in mind that users are no longer necessarily tied to a physical machine and/or location. Therefore, in some DV deployments, defining data access policies based on user, user type, location, and network attachment point might be required. These policies will impact how the DV infrastructure interacts with enterprise AD infrastructure.

Defining IT policies for user profiles that clearly specify use of DV-specific features such as USB redirection, localized printing, audio channel support, single sign on (SSO) and so on, per user based on the business case. Policies should also be defined per user profile or per user to allow/disallow use of

peripheral devices, changing monitor resolutions, choosing a specific display protocol, access to personal storage space, rebooting of HVD and so on. These features in a DV environment heavily impact network, storage, and compute requirements, and allowing all users access to these features can severely impact the viability of a DV deployment.

DV Endpoints

Selection of the user's endpoint is typically driven by application requirements and work related tasks. DV endpoint devices are used to access the user's HVD over the network, and come in one of three types: Zero-client, Thin-client, or Thick-client. Each type offers different features, has different data flows, and places different loads on the network. Also, based on the capabilities of the device, each device type is suited for a particular end user profile.

Zero Clients

Zero-clients are the simplest devices. They have embedded operating systems that are not exposed to the user. Zero-Clients have reduced local capabilities (e.g. reduced CPU, smaller memory footprint, and little to no local storage) and depend almost exclusively on the resources available within the HVD. Since there is no exposed operating system, there is no risk of virus infection, making Zero-Clients very secure. Zero-Clients typically rely on a single network connection to the HVD, making network policies for this class of device very simple. A simple Security Access Control List controlling the network flows admissible to/from the device can be created between the endpoint's network location (such as, subnet) and the connection broker (and associated virtual desktop pool). Some Zero-Clients can use 802.1x if Network Access Control (NAC) is desired as part of the security policy.

Thin Clients

Thin-client devices usually contain more local capabilities and often have a customizable locally embedded operating system (such as Linux or Windows embedded). Thin-Clients provide greater flexibility, as they can be customized and then locked down by the system administrator. The process of locking down a Thin-Client minimizes the risk of virus infection, and prevents the user from making local changes to the endpoint's configuration. Thin-Clients are typically used when applications local to the DV endpoint are needed; examples include web browsers, rich media content players, email clients and office automation tools running on the local operating system. The use of such local applications may shift the traffic patterns of Thin-Clients in comparison to Zero-Clients; this generally makes the network security policies more elaborate and may be dependent on the user's application selection and preferences. In some cases, media redirection allows for traffic to originate from, and terminate at the DV endpoint, without flowing through the HVD. For instance, if the user opens a local browser, the resulting traffic will not be routed through the user's HVD.

Thick Clients

Thick-Client devices refer to personal computers (including Laptops) running a standard operating system and relying on a locally installed DV software client to connect the device to the user's HVD. Thick-Client devices allow users to work offline and are often the choice of the "Mobile user". The traffic patterns for Thick-Clients may be very different than those of Thin-Clients. The user may choose to run most applications locally and only use a HVD to run specific applications in the Datacenter. For example the user can run the web-browser, email and office applications on the local Thick-Client, but

uses a HVD connection to access confidential documents that cannot reside on the local machine. Many options are available from Desktop Virtualization vendors for converting standard PCs and Laptops into Thick-Clients.

Performance and Capacity Planning

This chapter describes the tools and methods to use for capacity planning an end-to-end desktop virtualization deployment.

In an enterprise network, performance and capacity planning for desktop virtualization services has three fundamental dimensions:

- Capacity planning for compute and storage needs
- Capacity planning of infrastructure components necessary to support the services
- WAN capacity planning for delivering the service to remote sites

It is important to note that recommendations and results presented in this chapter are based on testing done across an end-to-end Cisco VXI system with optimized user profiles and a Cisco VXI workload. Readers can use this information to guide them in capacity planning but should carefully consider their own workloads and environments to make adjustments as needed for their own deployments. Also, this chapter is not a comprehensive guide for scaling every Cisco and third-party component used in the Cisco VXI system – for these, the reader is best served by reading the product documentation directly.

Computing and Storage Capacity Planning

When migrating from physical desktops to virtual desktops, it is important to have a good understanding of the user base and their resource requirements in terms of CPU, memory and disk space. Many enterprises transitioning to desktop virtualization also see it as an opportunity to do Windows 7 migration so understanding the implications of such a transition is also important.

A key step in the process is to group the users being migrated based on similar compute and storage needs such that capacity planning can be done at the group level. Similarly, disparate user groups should be estimated separately; otherwise, capacity may be wasted or be too constrained resulting in poor user experience. This type of grouping enables the administrator to develop a base profile for each user group based on common factors such as same operating system, applications and usage patterns. The base profile should also include resource utilization metrics such as CPU, memory, bandwidth and disk I/O utilization that is representative of the group by collecting data over a period of time and from a statistically significant number of users within that group. The base profile can then be used to estimate the compute and storage needs of the user group in a virtualized environment. Therefore, from an overall resource estimation and capacity planning perspective, administrators should group users with similar workloads and environment rather than across the entire user base being migrated – especially in large desktop virtualization deployments.

The next few sections describe the steps involved in planning for the compute and storage needs of an end-to-end virtual desktop deployment. The approach taken can be summarized as follows:

- Develop a base profile of the users in their physical desktop environment that includes a workload profile and associated resource utilization metrics such as CPU, memory, storage and bandwidth utilization
- Group users based on common factors (e.g. workload) that impact compute and storage needs
- Estimate the resource requirements of the user group targeted for desktop virtualization

- Estimate the per-server capacity in terms of the number of virtual desktops, including the associated storage capacity and performance required to support them. This estimate should take into account factors such as the effect of peak use, login storms, and outages and server downtime.
- Evaluate optimizations and other best practices that can improve resource utilization and performance in a production deployment
- Validate the resource estimations using a representative workload for that environment
- Extrapolate the single server data to determine the overall compute and storage hardware needs for the larger Cisco VXI deployment.

User Workload Profile

A base profile used in capacity planning should include the applications, the activities within those applications and a pattern of usage that is representative of the user group. This defines the group's workload profile that can then be used to classify the user group into one of the generic workload profiles commonly used by vendors to characterize users with similar compute, storage and networking needs. The profiles serve as the basis for performance and scalability data as load generation tools used for benchmarking use these profiles to generate a corresponding workload. The intensity of the workload can vary the scale and capacity significantly and as such it is key to any data used for estimating hardware resource needs. Workload profile is also critical if any testing is done to validate the resource estimations as the workload profiles used by load generation tools should closely match the user group's workload profile in order to accurately size the environment. In short, workload profiles play a critical role in the overall capacity of the desktop virtualization system being deployed. Generic workload profiles are described in greater detail later in this document.

Resource Utilization in the Current Environment

An important factor for estimating resource requirements in a virtualized environment is the resource utilization in the current physical desktop environment. Therefore, for a given target user group being migrated to virtual desktops, it is important to have a full understanding of the current environment and to characterize the resource utilization in terms of the following metrics:

- Average and peak CPU utilization
- Average and peak memory utilization
- Storage
  - Per-user storage capacity
  - I/O operations per second (IOPS)
  - Throughput (in bytes per second)
- Bandwidth utilization on the LAN

Administrators should monitor the use pattern of the target user group and determine the average and peak utilization for each of these in the environment. Monitoring should factor in potential variations in use pattern based on geographical location, when users log on for the day including shift transitions for environments that work in shifts, timing of backups, virus scans, and similar activities.

The resource utilization of the average physical desktop user can be determined as follows:

- CPU utilization: Use tools such as Microsoft Windows Perfmon or the Stratusphere tool from Liquidware Labs to collect average and peak CPU utilization from physical desktops in the target user group being migrated. Collect this data from a statistically significant number of desktops over a period of time while there is significant workload. You can then use a statistical average from the collected data to determine the peak and average CPU utilization for the group.
- Memory utilization: Also collect data about memory utilization on a physical desktop using the same tools as for CPU utilization, or a similar tool. As with CPU, analyze the data collected over a period of time from a significant number of desktops to determine the statistical averages of the group in terms of peak and average memory utilization. This data will be used to determine the memory needs of the group when using a virtualized desktop.
- Storage capacity and performance (IOPS and throughput) of the physical desktop: You can also determine IOPS and throughput from the physical desktop using the same tools as for collecting CPU and memory utilization information. Determine the peak and average data for the group in the same way as for CPU and memory utilization. This data will be used to determine the storage requirement of a virtualized desktop in that group.

Once the administrator has characterized the average and peak resource utilization for the group using a physical desktop, the process of estimating the compute, storage, and networking needs for migrating to a Cisco VXI system can begin.

Estimating Resource Requirements in a Virtualized Environment

To accurately estimate the resource requirements in a virtualized environment, several factors must be considered. In this section, we will take a closer look at three of these factors, namely CPU, memory, and storage. The data gathered in the previous section in terms of CPU, memory and storage can be used to estimate the number of virtual desktops a given server in the data center can support. Virtualization does introduce additional factors so the above resource requirements may need to be adjusted before estimating server capacity. Capacity of a single Server capacity can now be used to estimate hardware resources or servers necessary for a large-scale deployment.

Estimating CPU

To estimate the CPU resources needed in a virtualized environment, you can use the data from the physical desktops, as shown in the following example:

- Average CPU utilization for the physical desktops in the target user group is 5 percent.
- Assuming that the physical desktops are using a single-core 2.53-GHz processor, the average CPU requirement for the desktop is 5 percent of 2.53 GHz = 126.5 MHz.
- VMware recommends using a guard band of 10 to 25 percent to handle the following:
  - Virtualization overhead
  - Peak CPU utilization
  - Overhead associated with using a display protocol
  - Spikes in CPU utilization

  Therefore with a conservative guard band of 25 percent, the aggregate CPU requirement for a given desktop is approximately 158 MHz.

- The CPU requirement for a virtualized desktop, along with the computing capabilities of the blade server chosen for the deployment, can be used to estimate the number of virtualized desktops that can be supported on a given blade. For the Cisco UCS blade servers, the processing capabilities are listed in Table 1.

Note Each server model supports different processor types, though only one is shown per server in Table 1.

Table 1 UCS Server Model and Processor

Server Model	Processor
Cisco UCS B200 M1 Blade Server	Two quad core Intel Xeon 5570 series processors at 2.93GHz
Cisco UCS B200 M2 Blade Server	Two six core Intel Xeon 5680 series processors at 3.33GHz
Cisco UCS B250 M1 Extended Memory Blade Server	Two quad core Intel Xeon 5570 series processors at 2.93GHz
Cisco UCS B250 M2 Extended Memory Blade Server	Two six core Intel Xeon 5680 series processors at 3.33GHz
Cisco UCS B440 M1 High-Performance Blade Server	Four eight core Intel Xeon 7560 series processors at 2.26 GHz
Cisco UCS B230 M1 Extended Memory Blade Server	Two eight core Intel Xeon 7560 series processors at 2.26GHz

For additional information about the Cisco UCS server chassis and blade servers, please refer to the Data Center chapter.

- Using computing power as the only criterion, you can calculate the number of desktop virtual machines on a given blade server as shown here. For example, the number of virtual desktops that a Cisco UCS B250 M2 server can support would be:

```
Total compute power = 2 socket x 6 core a 3.33GHz = 39.96 GHz
Average CPU utilization of desktop = ~158MHz
Number of virtual desktops per server = 39.96GHz/158MHz = 252 desktops
```

Note Since this is strictly a theoretical exercise, please use actual data from testing when performing capacity planning for specific customer deployments.

Similarly, the number of virtual desktops that a Cisco UCS B200 M1 server can support would be:

```
Total compute power = 2 socket x 4 core a 2.93GHz = 23.44 GHz
Average CPU utilization of desktop = ~158MHz
Number of virtual desktops per server = 23.4GHz/158MHz = 148 desktops
```

Note This estimate is theoretical, based on a single factor (CPU). A number of other factors need to be considered to determine the actual number of virtual desktops that can be supported on a given server blade.

Note that the preceding estimation for the sizing a single server could be lower or higher, depending on the average utilization measured from the physical desktops. Similarly, the Cisco UCS server model chosen for a given desktop virtualization deployment can also affect the number due to differences in the computing capabilities of the server options available on the Cisco Unified Computing System™.

The number determined from the preceding sizing exercise could be the estimation used for estimating the overall server needs of a Cisco VXI deployment, if computing power is identified as the limiting factor. However, this scenario cannot be assumed to be the case until you have performed a similar exercise using memory and storage. Moreover, a number of other factors have to be considered before an estimation can be considered final for a given server blade.

Estimating Memory

To estimate the overall memory requirements in a virtualized environment, use the same methodology used for estimating CPU. The memory estimate for a single virtual desktop can be calculated from the statistical average determined from the physical desktops, as shown in the following example.

- Average memory utilization for the physical desktops in the target user group is approximately 750 MB.
- To accommodate additional memory demands due to spikes in memory utilization or additional applications, a 25 percent increase in the estimate is used. The aggregate memory requirement is therefore approximately 938 MB.
- The memory requirement for a virtualized desktop, along with the physical memory resource available on the blade server chosen for the deployment, can be used to estimate the number of virtualized desktops that can be supported on a given blade. The memory capacity on the various Cisco UCS server models is listed in Table 2:

Table 2 Cisco UCS Memory Capacity

Server Model	Memory Capacity
Cisco UCS B200 M1	96 GB
Cisco UCS B200 M2	96 GB
Cisco UCS B250 M1	384 GB
Cisco UCS B250 M2	384 GB
Cisco UCS B440 M1	256 GB
Cisco UCS B230 M1	256 GB

For additional information about the Cisco UCS server chassis and blade servers, please refer to the Data Center chapter in this design guide.

- Using memory as the single criteria for sizing the hardware needs in a Cisco VXI deployment, you can calculate the number of desktop virtual machines that a blade server can support as shown here. For example, the number of virtual desktops that a Cisco UCS B200 M1 server can support would be:

```
Memory Capacity = 96GB
Average memory requirement for a virtualized desktop = ~938 MB
Number of virtual desktops per server= 96G/938 M = 102 desktops
```

As with CPU estimation, the estimated number of virtual desktops on a single server may be lower or higher, depending on the data gathered from the physical desktops and the model of Cisco UCS server selected for the deployment. Also, this data can be used to extrapolate the total number of servers needed for a given Cisco VXI deployment if memory is determined to be the limiting factor.

Note that the memory utilization from a physical desktop used in the preceding calculation can vary depending on the guest OS and applications deployed in a given environment. An alternative but also a less accurate method of estimating the number of virtual desktops is to use the minimum recommendation from Microsoft for per–virtual machine memory utilization, as shown in Table 3. However, for Microsoft Windows XP, the minimum recommendation shown here should be increased to 512 MB to accommodate Microsoft Office applications and other applications that may be running on the desktop. The memory configuration used for virtual desktops in an end-to-end Cisco VXI system is also provided as an example. The memory configuration was sufficient to provide a good user experience for the workload profile validated in the end-to-end system.

Table 3 Memory Configuration

Microsoft Windows OS	Minimum (Microsoft) Memory Requirement	Memory Configuration in Cisco VXI system using Knowledge Worker + Profile
Microsoft Windows XP with Service Pack 3	256 MB	1 GB
Microsoft Windows 7 32b	1 GB	1.5 GB
Microsoft Windows 7 64b	2 GB	2 GB

Estimating Storage

For storage, the average IOPS and throughput data collected from monitoring the physical desktops can be used as the storage requirements for the virtualized desktops. For example, if the average IOPS is 5 and the average throughput is 115 kbps, then the same IOPS and throughput values should be expected when the desktop is virtualized. For a desktop virtualization deployment, the factors summarized here can also have a significant effect and should be considered when sizing storage needs. For example, IOPS can peak when:

- Users are powering on: When users come in at the beginning of the workday and start powering on their virtual desktops, IOPS and throughput will peak, a situation referred to as a boot storm.
- Users are logging on: Though the virtual desktops do not need to be powered on, there can be peaks in storage I/O as users are logging on in the morning to start their work. This situation is referred to as a login storm.

- Other activities occur: Activities such as antivirus scan and backups can cause storage performance requirements to spike.

Some applications specific to a customer environment can cause similar spikes in storage I/O. All these factors must be taken into account when designing the storage environment for Cisco VXI deployments.

Another aspect that needs to be considered when sizing storage needs is the disk space allocated to a virtual desktop. You can calculate this space by adding the storage requirements required for each of the following items:

- Operating system and base set of applications
- Page and swap files and temporary files created by the OS and applications
- Page and swap files created by a VMware ESX and ESXi host for every virtual machine deployed on the host (equals the memory allocated for the virtual machine)
- Microsoft Windows profile (user settings such as desktop wallpaper)
- User data (equivalent to the My Documents folder in Microsoft Windows)

For an example of the storage allocation to use for virtual desktop machines, see Table 4. Note that deploying desktop virtualization using View linked clones or Citrix streamed desktops with a provisioning server will minimize the per-desktop disk space necessary for windows and applications as large groups of desktop pools can share the same master virtual machine image. As a result, only the delta between that and the available disk space on the master VM may need to be allocated on a per-desktop basis by using deployment models mentioned above. However this delta disk size can also be minimized by refreshing the OS disk periodically or by using non-persistent desktops – so a number of options exist in this regard as well.

Table 4 Storage Allocation for Desktop VM

Guest OS on virtual desktop	Minimum Disk Space Windows and Applications	Windows Page File and Temporary Files	Hypervisor Swap File	Windows User Profiles	User Data
Windows XP	10 GB	3 GB	1 GB	2 GB	5 GB
Windows 7 (32-bit)	16 GB	4 GB	1.5 GB	2 GB	5 GB
Windows 7 (64-bit)	20 GB	4 GB	2 GB	2 GB	5 GB

Estimating Server Capacity

As stated before, several factors can influence the performance and scalability of a server. The estimation for the number of virtual desktops on a given server can yield a different number if each factor is considered independently. For this reason, the estimations performed in the Estimating CPU and Estimating Memory sections earlier in this chapter for a Cisco UCS B200 M1 server are theoretical exercises. However, the data (summarized in Table 5) can aid in finding the limiting factor for a given server, as well as provide the initial virtual machine density to target if testing is performed to validate the estimation using the specific workload for that environment.

Table 5 Estimated Capacity

Factor Used to Determine Capacity	Average Value for a Virtualized Desktop	Server Capacity (Theoretical for Cisco UCS B200 M1)
CPU	158 MHz	148
Memory	938 MB	102

Note The estimates in Table 5 are not the actual capacity of the server. They are **theoretical** estimations based on the CPU and memory utilization assumptions for the user group in a given environment.

In any desktop virtualization deployment, workload is one of the most critical factors for accurately estimating the virtual desktop density on a given server in that environment. Therefore, you must use a workload that closely matches the user group's workload for any estimation and testing you perform to determine the per-server sizing. See link below for actual results from a scalability testing done with XenDesktop:

http://www.cisco.com/en/US/solutions/ns340/ns414/ns742/ns743/ns993/landing_dcVirt-vdi.html

In addition to the above scalability stud, single server scalability testing was also done across the end-to-end Cisco VXI system discussed in this design guide. Testing was done for a number of deployment profiles across the end-to-end Cisco VXI system similar to how customers would deploy and use the system. The testing was done using test tools located in the campus network that initiate VDI sessions to the virtual desktops hosted in the data center where the session would span the following:

- Campus network that consists of access, distribution and core network layers built using catalyst 3500, 4500 and 6500 series layer 2 and layer3 switches
- Data center network, also with a core, aggregation and access layer made up of Nexus 5000 and Nexus 7000 series built in accordance with Cisco validated data center infrastructure design
- Data center Services aggregation layer where firewalls are used to control all traffic entering and leaving the data center. Firewalling is also used at the access layer within the data center to control traffic between virtual desktops from all other services and application infrastructure residing in the data center. Redundant ACEs are used for load balancing all connections to application servers used for providing desktop virtualization services such as Citrix Web server and XenDesktop controllers
- Data Center that consists UCS 5108 B-series chassis with B250 M2 and B200 M2 blade servers connected to Nexus 1000 series access switches. Both NAS and SAN based storage were used depending on the needs of the deployment profiles tested.

For details on the results and data from the single server testing done in the Cisco VXI system using a Cisco VXI workload – please see the Workload Considerations section of this chapter.

In addition to the workload, there are a number of other factors apart from CPU, memory, and storage utilization can influence a server's scale and capacity numbers and are discussed in the next few sections.

Hypervisor Considerations

Virtual CPU

Virtual CPU requirements may vary depending on the desktop virtualization solution being deployed. Citrix recommends deploying 2 vCPUs with HDX though 1vCPU is also common in desktop virtualization environments with a wide range of workloads. If more than one vCPU is used and if the hypervisor is ESXi, VMware recommends not mixing desktops with 1 and 2 vCPUs on a single server due as the resource scheduling can result in 1vCPU machines from being serviced before the 2vCPU desktops. Though resource pools can be used to reserve CPU shares can be performed, these are generally not used in virtualized desktop environments. VMware has increased the number of vCPUs that can be supported on a given host to the numbers shown in Table 6, and these values are worth noting from a capacity planning perspective. However, for servers capable of supporting 25 vCPUs per core with a high number of cores, the limiting factor for the number of virtual desktops can be the number of virtual machines supported per host, as the table shows. The data in Table 6 is from VMware's configuration maximums document for vSphere 4.0 and 4.1.

Table 6 Virtual CPU Limits

Limits	VMware vSphere 4.0	VMware vSphere 4.1
Virtual CPUs per core	25	25
Virtual CPUs per host	512	512
Virtual Machines per host	320	320

Note Though the number of vCPUs supported per core is 25, the number of achievable CPUs per core depends on the workload.

Note In actuality, the number is probably closer to the 8 to 10 virtual desktop virtual machines per core in desktop virtualization deployments.

Memory Considerations

The transparent page sharing (TPS) feature available on ESXi hypervisor can significantly reduce the memory footprint, particularly in desktop virtualization deployments where the OS and applications data may have a lot in common across different desktop virtual machines. TPS uses a background process to monitor the contents of memory and evaluates the data being loaded to determine if it is the same as what is already in memory. If it is the same, the virtual machine attempting to load the duplicate data will be redirected to existing content in memory, thereby enabling memory sharing. TPS can be thought of as memory de-duplication feature and is enabled by default. For more information about this feature, see

http://www.vmware.com/resources/techresources/531

In a Cisco VXI environment, transparent memory sharing enables a server to accommodate a larger number of virtual desktops on a single blade, at least from a memory perspective though this may not be a limiting factor.

Since TPS uses redundancy to share and overcommit memory between virtual machines running on a host, the workloads on these virtual machines should be as similar as possible. To optimize the effect of TPS in a Cisco VXI environment, you should group virtualized desktops of users with similar workloads, such as the same guest OS (Microsoft Windows 7 and Windows XP) and applications (Microsoft Office and antivirus applications), on the same host to optimize the effect of TPS.

TPS behaves differently on the newer hardware-assisted virtualization processors, such as the Intel Nehalem and Westmere processors that are used on Cisco UCS servers. The newer processors use memory pages that are 9KB in size and improve performance by 10 to 20 percent. TPS operates on 4-KB pages to eliminate duplicate data. With these newer processors, TPS is not in effect until the available memory reaches a minimum and there is a need to overcommit memory. A background process is still monitoring and scanning the memory pages to determine when TPS takes effect. See the following VMware Knowledge Base articles for more information about TPS with the newer hardware assisted virtualization processors:

- TPS in Hardware Memory Management Unit (MMU) Systems: http://kb.vmware.com/kb/1021095
- TPS Is Not Utilized Under Normal Workloads on Intel Xeon 5500 Series CPUs: http://kb.vmware.com/kb/1020524

Also note that VMware studies have shown that TPS does not have any effect on the performance of the host and therefore recommends the use of this feature. Please contact VMware for additional information about TPS.

High Availability

For a desktop virtualization deployment of significant scale, high availability of the virtual desktop is a primary concern for most deployments. For example, VMware recommends deploying hosts or servers in clusters to enable high-availability features such as VMware High Availability (HA), Dynamic Resource Scheduling (DRS), Fault Tolerance, and vMotion. With clustering, server/compute resources can be pooled together such that DRS can be deployed to dynamically load-balance virtual desktops across hosts in the cluster based on resource needs or configured to do so only at startup.

However, deploying servers in a cluster can change and potentially limit the maximums that VMware supports. The supported limits are available through VMware configuration maximums and are available when new releases change the supported limits. Table 7 lists some of the data relevant to sizing a desktop virtualization deployment. This table should be reviewed for planning any large-scale deployment of Cisco VXI. For a complete set of configuration maximums, refer to your hypervisor documentation.

Table 7 Configurations Maximums

Limits	VMware vSphere 4.0 Update 2	VMware vSphere 4.1
Number of virtual machines per host	320	320
Number of vCPUs per host	25	25
Hosts per high-availability cluster	32	32
Number of virtual machines per cluster	-	3000
Number of virtual machines per host with 8 or fewer in the high-availability cluster	160	-
Number of virtual machines per host with more than 8 hosts in a high-availability cluster	40	

Limits	VMware vSphere 4.0 Update 2	VMware vSphere 4.1
Number of hosts per VMware vCenter server	1000	1000
Number of hosts per data center	100	400

Power Management

The power management policy used in a desktop virtualization environment can have an impact on the compute resources. Virtualization vendors often recommend that virtual desktops be put in a suspended state when not in use. The suspended state is an optimal configuration that enhances the user experience while reducing resource (CPU and memory) use. If all virtual machines are left powered on, the host resources cannot be used by other virtual machines on the same server. With persistent desktops, the virtual machine can be immediately suspended when the user logs off.

Storage Considerations

Desktop virtualization solutions generally reduce overall storage needs by enabling pooled desktops to be deployed with a main OS disk that is shared among all desktops in the pool. Using pooled desktops greatly reduces the aggregate storage capacity necessary for migrating to a virtualized environment. Although the cost of shared storage is significantly higher than separate disks on laptops and desktops, virtualization reduces overall storage costs due to the ability to share the same OS disks among many desktop virtual machines.

Operating System Disk

In desktop virtualization deployments, the OS disk refers to the parent virtual machine's virtual disk on which the guest OS (Microsoft Windows XP or Windows 7) and applications (Microsoft Office) are installed. This OS disk is read by all desktops in the pool, resulting in significant storage savings, since a single OS disk can be used by a large number of desktops without each having to maintain its own OS disk. Ideally, this disk should be read-only for both storage and operation efficiency, but it can be used as a read-write disk to store the following types of typical Microsoft Windows desktop data:

- Microsoft Windows profile data
- Temporary files, including page files
- User data

For better storage and operation efficiency, the OS disk should be kept as a read-only disk, and the data listed here should be redirected to another location as follows.

- Microsoft Windows profile data can also be redirected to a Microsoft Windows share or to another virtual disk dedicated for this purpose, so that the data can be saved in the event that the OS disk is updated.
- Temporary files can also be redirected to a non-persistent disk so that the data can be flushed to reduce storage use. A separate location on the SAN or on a transient volume on network-attached storage (NAS) can be used.
- User data that is typically saved in the My Documents folder should be redirected to a Microsoft Windows share or to a separate disk.

Thin Provisioning

Thin provisioning is a way to conserve storage resources and increase storage utilization in a virtualized environment. With thick provisioning, when a virtual machine is deployed, the virtual disk associated with the virtual machine is given its full allocation of storage regardless of whether it uses it, resulting in wasted space. With thin provisioning, this inefficiency is reduced by allocating storage resources only when the virtual machine needs them. Therefore, a virtual desktop running Microsoft Windows 7 with a 20-GB disk will not have 20 GB of disk space reserved on the storage system (SAN or NAS), though Microsoft Windows and applications running on the desktop will operate as if it has the full 20 GB of space allocated to it. Therefore, thin provisioning enables the efficient use of the underlying storage resources and improves the scalability of the aggregate storage capacity by over-committing the storage. In a DV environment, this approach results in a higher number of virtual desktops that can be supported with the given storage capacity.

Thin provisioning can be done by the hypervisor or by the storage systems. Storage vendors such as NetApp and EMC offer thin provisioning at the storage level that further improves any storage efficiency enabled by Hypervisor level thin provisioning. With storage thin provisioning, the actual state of the storage allocation is hidden from the Hypervisor by the storage system.

Note Since thin provisioning is an over allocation of the storage resources, you should carefully monitor the state of the thin-provisioned disk so that additional storage can be added to the data store before a lack of space causes problems.

Storage Footprint Reduction

Storage vendors support technologies that offer economies of scale for sizing the storage needs of a desktop virtualization deployment. Technologies include data de-duplication to increase storage efficiency by eliminating redundant information in the data being stored. This feature can be used for primary file systems and end-user file data in VMware and other virtualized environments. If the duplicate data is from different virtualized desktop virtual machines, the data is stored only once and the metadata associated with it is changed so that both virtual machines have access to the data. As with thin provisioning, de-duplication can provide significant storage efficiencies, improving desktop virtualization scalability since the existing storage can now support a larger number of desktop virtualization desktops. Therefore, enabling de-duplication, particularly in large desktop virtualization deployments, is highly recommended.

Please refer to EMC and NetApp documentation for more information about using de-duplication in a desktop virtualization environment.

Partition Alignment

Microsoft Windows file system partitions running on virtualized desktops should be aligned with the underlying storage partitions. This alignment can improve storage performance by reducing overall I/O latency while increasing storage throughput. This alignment is currently needed only with Microsoft Windows XP because Microsoft Windows 7 (32-bit and 64-bit) automatically provides this alignment. The problem occurs because Microsoft Windows writes 63 blocks of metadata directly at the beginning of the drive, resulting in misalignment of the first partition created on the disk. As a result, the drives may need to read an extra block of data unnecessarily, causing additional IOPS on the drive. To address the misalignment problem, an aligned partition is created on the drive that aligns with the storage system used. Both block-based and file-based storage systems can benefit from this alignment. In the Cisco VXI

system, a 64-KB aligned partition was created on the parent virtual machine of the desktop pools to align with EMC's SAN and NetApp's NAS storage. Please refer to Microsoft and Citrix documentation for information regarding this.

Storage Network

- Jumbo frames: Cisco VXI deployments using IP-based storage should enable jumbo frames to increase storage bandwidth utilization and improve I/O response times. Jumbo frames increase the maximum transmission unit (MTU) for Ethernet frames used to transport IP traffic in data center LAN networks. Enabling jumbo frames increases the Ethernet MTU from the default value of 1518 bytes to 9000 bytes typically and should be enabled on every link between the server hosting the virtual desktops and the IP storage it uses. Jumbo frames not only improve overall throughput, but also reduce the CPU burden on the host for large file transfers.
- Separation of storage network: Storage traffic should be physically (ideal) or logically separated from other network traffic using VLANs. Cisco Unified Computing System architecture supports two host bus adapters (HBAs) dedicated to storage if Fibre Channel–attached SAN storage is used.
- Similar physical separation is recommended for IP-based storage, such as storage of NFS and Small Computer System over IP (iSCSI) traffic. A separate VMkernel port and VLAN should be used for IP storage traffic using a dedicated uplink port on the host. This type of physical separation of the IP storage traffic from other IP traffic is possible using the latest converged network adapters (CNAs) on the Cisco Unified Computing System, which support 128 virtual uplink ports. If an uplink cannot be dedicated, the separate VLAN used for storage will provide the logical isolation. The physical isolation should be extended into the data center network by using dedicated ports or switches at the access layer where Cisco Nexus 5000 Series Switches are typically deployed. If the storage traffic extends into the aggregation layer of the data center network, Cisco Nexus 7000 Series Switches that are typically deployed at this layer support physical separation through the virtual data center (VDC).
- PortChannels: To increase the aggregate uplink bandwidth without sacrificing availability, PortChannels can be used between the LAN uplink ports on the host and the access layer switch (Cisco Nexus 5000 Series). This approach is important if the Cisco VXI deployment uses IP-based storage since it significantly increases the LAN bandwidth required.
- Multipathing: For both block-based SAN storage and IP-based NAS storage, multipathing can be used to create load-balanced but redundant paths between the host and the storage it uses. VMware learns the various physical paths associated with the storage device, and it uses a path selection scheme to determine the path a given I/O request should take. The three options on VMware for selecting the path to the storage device are fixed, most recently used, and round-robin. Round-robin should be used to load balance I/O traffic across multiple physical paths. Because of the performance improvements the multipathing provides through enhanced storage resiliency, Cisco VXI deployments should enable multipathing if the storage vendors support it. Both EMC (Fibre Channel SAN) and NetApp (NFS), included in the end-to-end Cisco VXI system, support this capability.

Guest OS Optimizations

Microsoft Windows can be optimized to improve the performance and scalability of virtualized desktops as outlined below.

- Optimize the Microsoft Windows virtual machine file system for optimal I/O performance by disabling the last-access-time updates process in NTFS. Microsoft Windows will update files with the last access update time when an application opens that file, and disabling this option will reduce the IOPS occurring within the file system.
- Enable Windows Best Performance especially for Windows 7 deployments it can provide performance important both from a compute perspective and from a WAN BW utilization perspective
- Prevent antivirus software from scanning the main OS disk that each virtual machine uses since it is deployed as a read-only disk with antivirus checks run against it before it was deemed as the golden master for use by the Citrix XenDesktop pool of virtual desktops. This step can help increase storage performance particularly in a large Cisco VXI deployment.

A number of windows optimizations can be enabled on virtualized desktops to improve performance. Table 8 shows some of the main optimizations implemented on the Cisco VXI system for Microsoft Windows XP and Windows 7.

Table 8 Microsoft Windows Guests OS Optimizations

Microsoft Windows Guest OS Optimizations
Microsoft Windows Guest OS Optimizations
Disable Microsoft Windows Hibernation
Disable Microsoft Windows Defender (N/A on XP)
Disable Microsoft Feeds synchronization
Disable Microsoft Windows Scheduled Disk Fragmentation (N/A on XP)
Disable Microsoft Windows Registry Backup (N/A to XP)
For PCoIP, set the power options for Display to off
Disable mouse pointer shadow
McAfee Anti-virus scan on write only
Disable Pre-fetch/Superfetch Service (N/A on XP)
Disable Microsoft Windows Diagnostic Policy Service (N/A to XP)
Enable “No automatic updates"
Disable System Restore (since refresh can be done by composer)
Disable paging of the Microsoft Windows OS itself
Disable unwanted services
Turn off unnecessary sounds at startup and shutdown
Disable indexing services
Delete all background wallpapers
Disable screen saver

Validating Capacity Estimates

The next step in the overall resource planning process is to validate the capacity estimations based on factors such as CPU, memory, and storage as well as factors outlined in the previous section. On the basis of the baseline performance data from the physical desktop and the theoretical estimation for the number of virtual desktops that can be rolled out on the Cisco UCS blade server chosen for the deployment, perform performance characterization and validation in a virtualized environment to determine the following:

- Average CPU utilization of the server
- Memory utilization of the server
- Storage IOPS generated by the server
- Network bandwidth utilization of the server
- Application response times

Workload Considerations

To validate capacity estimates for a Cisco VXI deployment, one option is to roll out the service to a pilot group and validate the resource estimation with actual users. Alternatively, the data regarding user activities, applications used, and use patterns collected from the physical desktops can be used to define a workload specific to that environment. For testing, the custom workload can then be automated to simulate the user workload, or it can be mapped to one of the generic profiles commonly used by workload generation tools used in scale and performance testing.

The workload defines the applications that a person actively uses, such as word processing, presentation, and other office applications, but it can also include background activities such as backups and antivirus scans. Workloads for users within a company will vary depending on a person's job or functional role and may differ according to the organization structure (sales, marketing, manufacturing, etc.). Workloads can also vary based on the time of day, particularly if the desktop virtualization users are geographically dispersed. Background activities that begin at specific times, such as backups and antivirus scans, can also increase workloads.

For a comprehensive look at workload definition and other considerations in an enterprise network, please refer to the Desktop Virtualization chapter.

Table 9 User Profiles

User Profile	Description
Task Worker	• One application open at a time • Limited printing • Limited mouse usage • Primarily text editing.

User Profile	Description
Knowledge Worker	• Multiple applications open at a time • Variety of applications • Graphical applications with multimedia and use of USB peripherals.
Power User	• Multiple applications open at a time • Graphical and/or computational intensive applications • User may need administrative rights.

For the end-to-end Cisco VXI system outlined in this document, a variation of the first two user profiles was used for testing and will be referred to as the Cisco VXI Knowledge Worker+ (KW+) profile. The details of this profile are outlined in the Table 10 below.

Table 10 Cisco VXI Knowledge Worker + Profile

Script Sequence	Applications in the workload profile	Activities within an Application
Step 1	Start Cisco Unified Personal Communicator	1. Check if Cisco Unified Personal Communicator application is installed 2. If not, move to the next application 3. If yes, launch Cisco Unified Personal Communicator 4. Log into Cisco Unified Personal Communicator
Step 2	Start Internet Explorer 8 and keep it running	1. Start application 2. Clear cache 3. Open first web page http://10.0.128.150/VXI/index.html 4. Open second web page http://10.0.128.150/VXI/Recipes/index.htm
Step 3	Start Microsoft Word and close it	1. Start application 2. Open an existing file 3. Navigate to last page 4. Insert a page 5. Write a paragraph 6. Save Word document 7. Close the application

Script Sequence	Applications in the workload profile	Activities within an Application
Step 4	Start Microsoft Outlook and close it	1. Start application 2. Wait until all folders are up to date 3. Perform Send/Receive 4. Clean Inbox by deleting existing e-mails 5. Send e-mail 6. Wait to receive back e-mail 7. Delete e-mail
Step 5	Start Microsoft Excel and close it	1. Start application 2. Open an existing file 3. Set zoom level to 100% 4. PageUp 5. PageDown 6. Save document 7. Close the application
Step 6	Start Microsoft PowerPoint and close it	1. Start application 2. Open an existing file 3. Play slideshow 4. Close the application
Step 7	Start Adobe Acrobat and close it	1. Start application 2. Open an existing file 3. Navigate to page 50 4. Set zoom level to 75% 5. Zoom up (5 times) and zoom down (5 times) 6. Close the application

- In Cisco VXI, McAfee MoveAV 1.5 is running in the background with a default scan profile
- All applications except for Cisco Unified Personal Communicator and Outlook are randomized across multiple iterations of the workload loop
- A random timer is used in each step above that pauses between 7 and 11 seconds to simulate user "think" time

In the following sections we look at the results of the single server characterization done for various Citrix deployment profiles, followed by application characterization done for Cisco collaboration applications in a Cisco VXI environment.

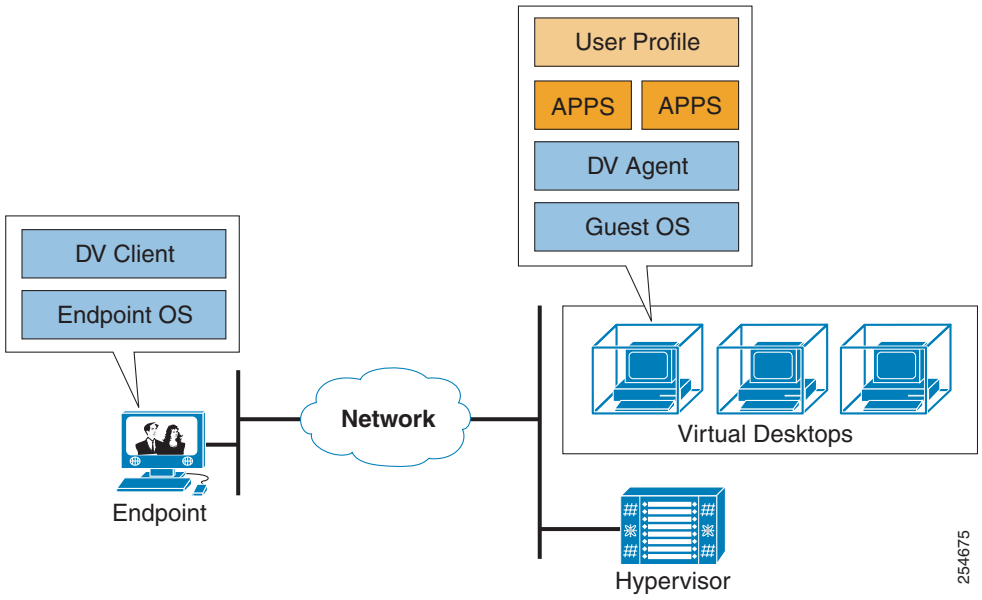
Managing Virtual Desktops

In a desktop virtualization environment, virtual desktops are organized in pools that are associated with end users. The association of a user with a pool allows the user to access the desktops in that pool. A dedicated pool consists of a single desktop with a single associated user, similar to a personal desktop. A shared pool includes multiple desktops with multiple associated users. In a shared pool, the assignment of desktops to users can be accomplished in several ways.

The user assignment can be performed statically through administrative configuration, or dynamically using the connection manager. When the connection manager is used, assignments can be made in a random or sticky pattern, so the user is always assigned the same desktop, or in a floating manner, with the user choosing a desktop. The desktops in a pool can be generated statically or dynamically (on demand), and they can be persistent (retained after the user logs off) or non-persistent (deleted after the user logs off.) A desktop can be newly created, or it can be cloned from a master desktop template.

From a desktop management perspective, you need to consider a few points when managing the Cisco VXI system (Figure 8). Generally, virtual desktops run the same OS as physical desktops; thus, the same component management strategy should be used for the OS, application updates, Microsoft Windows user profiles, and network settings. Applications can be installed locally, on the virtual desktop, or streamed from an application server. When managing persistent desktops (those retained after use), you must update each virtual desktop individually. When managing non-persistent desktops (those that are deleted after use) that are generated on demand using a master template, you should update only the master; subsequent desktops generated to service new user sessions will be cloned from the updated master template.

Figure 8 Desktop Management



To understand desktop management, consider the following scenarios:

- In one type of deployment, a personal desktop is created once, statically, by the administrator and assigned to a user for regular use. The user is allowed to customize settings and install software.

- In another deployment, a shared desktop is created dynamically (on demand) and assigned to a user who is a task worker. Even though the user is allowed to customize settings and install software, the changes made by the task worker are not permanent since the desktops are deleted after the user logs off. New users will be provided with new desktops, generated on demand.

There are advantages to managing virtual desktops. The task of creating a new desktop—for instance, installing the operating system and applications and customizing user settings for a new user or a user who requires an upgrade—is much faster when the desktop is cloned from a master virtual machine template that already has the OS, application, and user settings configured.

Applying updates to dedicated desktops (persistent) should be accomplished in a manner similar to that used for physical desktops. In this scenario, the administrator still maintains some control, but the user is given more flexibility as to applications installed and user settings (personalization). The tasks of maintaining software updates, downloading and installing new software, performing file backup, and running antivirus software are largely the same for virtual desktops as for physical desktops.

Applying updates to shared desktops (non-persistent) can be accomplished in a centralized manner because only the master template needs to be modified. Now with Cisco VXI deployments, IT can manage single instances of each OS, application, and user profile to greatly simplify desktop management. The administrator has complete control over the image version and application revisions on the desktops. Any changes and updates to master templates can be implemented by the administrator and made available to new user sessions immediately. This process can be completed either by statically cloning a virtual machine template or by using the VMware linked-clone or Citrix Provisioning Services feature for dynamically generated desktops.

Although some applications, such as antivirus software, can be installed locally on the virtual desktop, others, such as Microsoft Office, can be streamed to the virtual desktop from a central repository. When updating applications that are streamed, you need to update only the master copy on the central application server. When updating applications that are installed locally on the virtual desktop, you need to update the desktop or master template desktop.

Since virtual desktops reside in the data center and may not be in close proximity to network printers, the appropriate changes should be made to the applications that install and configure printers on desktops. Desktop virtualization vendor solutions support connectivity to USB devices that are attached to endpoints; thus, printing to a local printer connected through a USB port should be enabled for the end-user session.

Cisco Management Tools

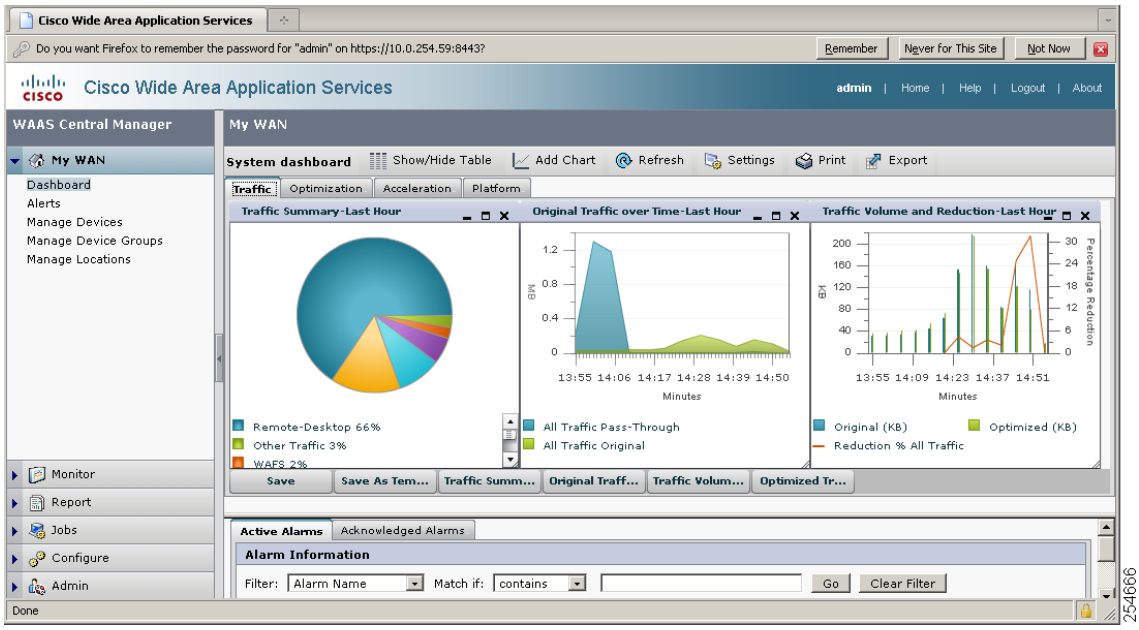
Cisco tools critical to management of the Cisco VXI system are Cisco UCS Manager, Cisco Fabric Manager, Cisco DCNM, CiscoWorks LMS, Cisco ACE Device Manager, Cisco WAAS Central Manager, and Cisco Adaptive Security Device Manager (ASDM). Many of these tools, such as Cisco UCS Manager, Cisco ACE Device Manager, Cisco WAAS Central Manager, and Cisco ASDM, are embedded in the managed device and are accessed through the browser. Some of these tools, such as CiscoWorks LMS, Cisco Fabric Manager, and Cisco DCNM, are Microsoft Windows server applications and are accessed through the browser or client applications.

Note Cisco WAAS Central Manager runs on a dedicated Cisco WAAS appliance.

Use the Cisco WAAS Central Manager to monitor and generate reports about traffic optimization for sessions initiated from branch offices across the WAN (Figure 9). Since the Cisco WAAS appliance can be deployed inline or offline, the traffic utilization and optimization reports generated by the central manager may not represent all the traffic flowing through the neighboring router. The Cisco WAAS

device will be aware only of traffic presented to it in an offline configuration. You should use a backup Cisco WAAS Central Manager with stateful redundancy to increase availability, and you should locate the Cisco WAAS Central Manager in the data center with other critical resources. After a Cisco WAAS configuration is defined on the Cisco WAAS Central Manager, it can be pushed out to all Cisco WAAS devices in the network in an efficient manner.

Figure 9 Cisco WAAS Central Manager GUI

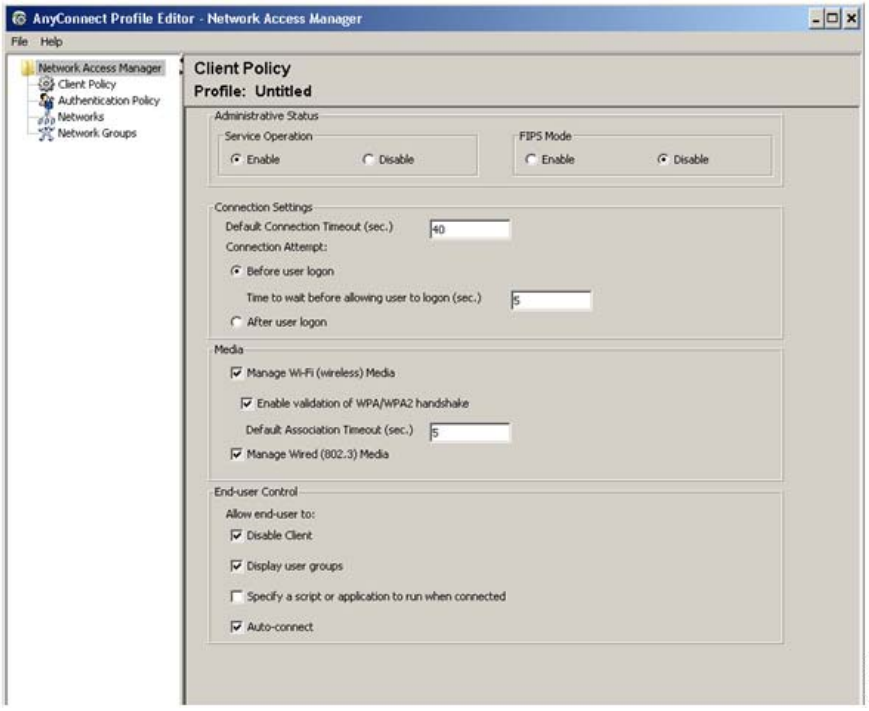


Use Cisco ASDM to monitor and generate reports about sessions initiated by teleworkers (located outside the enterprise) who use Cisco ASA to access the corporate network using a VPN client. Cisco ASDM can be used to provision firewalls, Network Address Translation (NAT), and VPN profiles on the Cisco ASA appliance as well as to monitor firewall activity, CPU use, network interface use, and VPN status on the Cisco ASA (Figure 10).

Figure 10 Cisco ASDM Device Dashboard

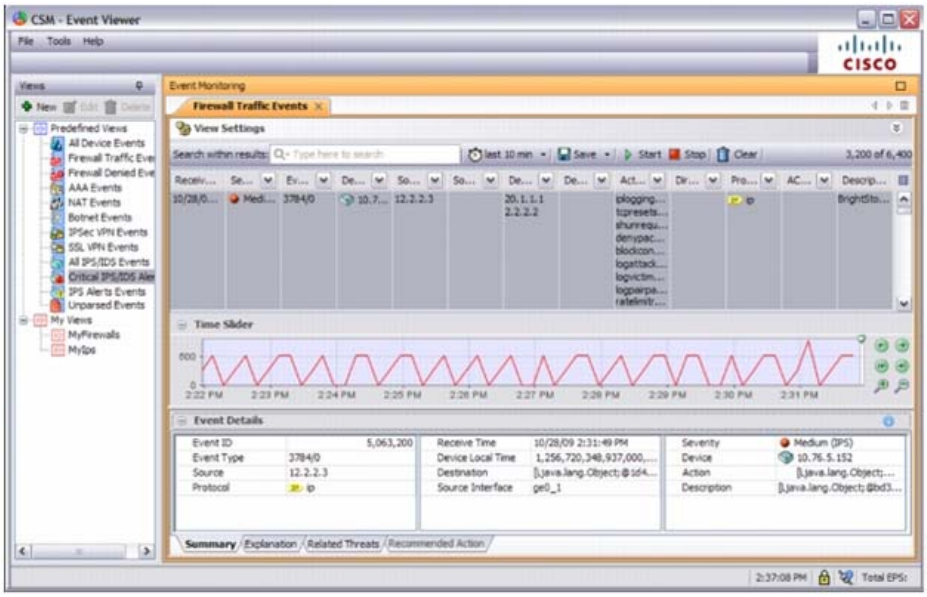
The AnyConnect profile editor is a GUI-based configuration tool that can be used to configure the AnyConnect client profile which is an XML file containing settings that control client features. Previously, profile settings could only be changed manually by editing the XML tags in the profile. The profile editor can also be launched from ASDM if the client software package is loaded on the ASA as an SSL VPN client image. It allows administrators to create configurations for AnyConnect VPN Client, Network Access Manager (Layer 2 connection) client, Web Security client and Telemetry client. The configurations ACPE generates can be bundled with AnyConnect binary installation package and pushed to endpoints via SMS.

Figure 11 AnyConnect Profile Editor



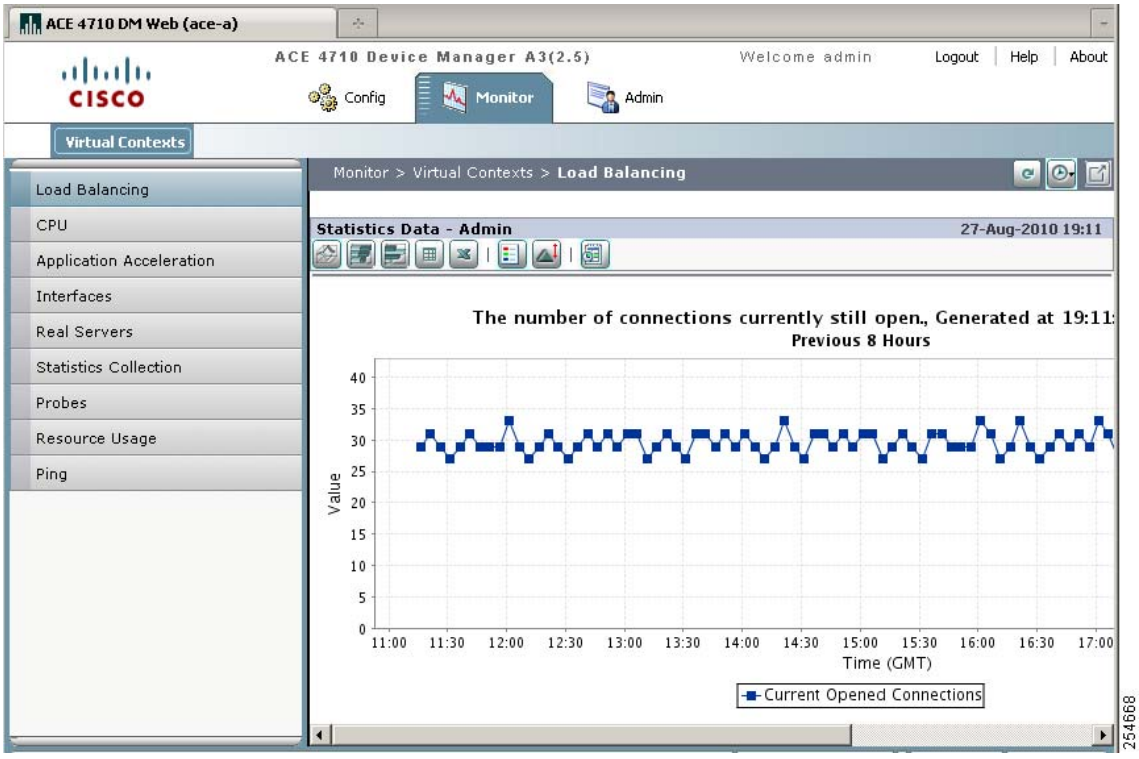
Cisco Security Manager manages security policies on Cisco security devices. It supports integrated provisioning of firewall, IPS, and VPN (site-to-site, remote access and SSL) services across IOS routers, Catalyst switches, ASA and PIX security appliances, Catalyst services modules related to firewall, VPN, and IPS, IPS appliances and various service modules for routers and ASA devices. It can be used to provision the ASA and DMVPN on ISR routers in a Cisco VXI deployment.

Figure 12 Cisco Security Manager Console



Use the Cisco ACE Device Manager to monitor and generate reports about all desktop virtualization user sessions that use a Cisco ACE appliance to access the connection manager. The Cisco ACE Device Manager can generate reports about the overall session load presented to the connection managers located in the data center (Figure 13). You should deploy Cisco ACE in redundant pairs with stateful redundancy to increase availability.

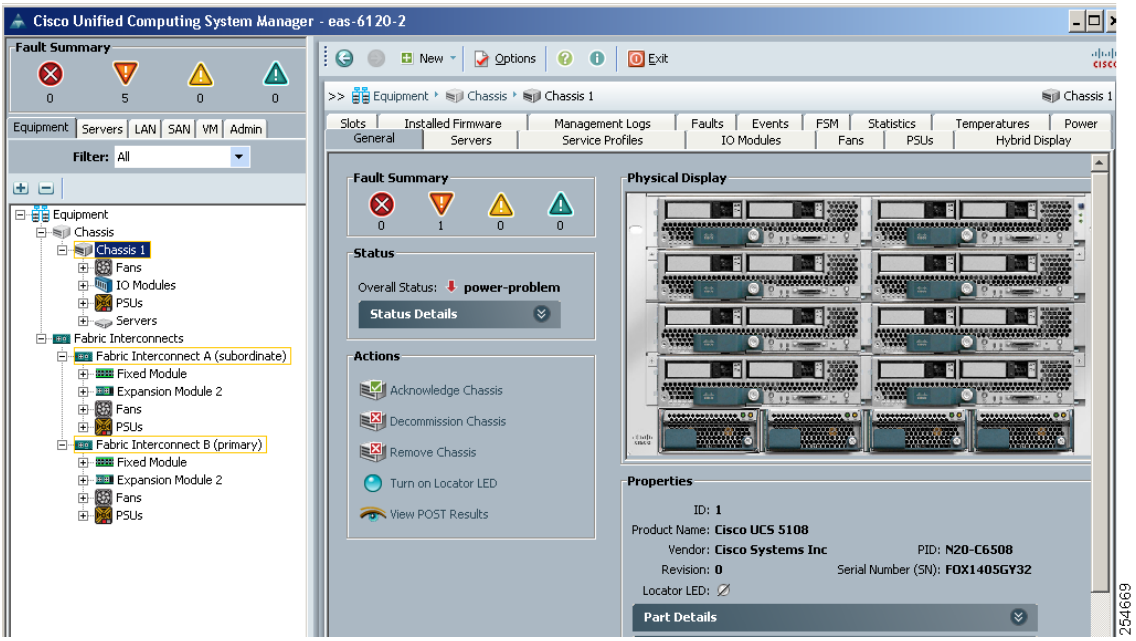
Figure 13 ACE Device Manager Monitor



You should use the web-based GUI (instead of the CLI) to manage Cisco WAAS and Cisco ASA, since the GUI can manage multiple devices, provision them with similar settings, and generate aggregate reports based on statistics collected from multiple managed devices. This approach is recommended for large-scale deployments, which have many devices to be managed.

Use Cisco UCS Manager (Figure 14) for provisioning a Cisco UCS blade server (both the Cisco UCS 5100 Series Blade Server Chassis and the Cisco UCS 6100 Series Fabric Interconnects) because it can scale up to 40 blades across multiple chassis. Provision separate VLANs on the uplink to separate management traffic from virtual machine traffic. Perform this provisioning on the Ethernet uplink ports and the vSwitch running in VMware ESXi. You should use the service profile templates, as well as templates for vNIC and virtual host bus adapter (vHBA) characteristics, to apply uniform configurations across all blades. Cisco UCS Manager also allows blade swapping to quickly reduce downtime during a hardware exchange.

Figure 14 Cisco UCS Manager



Use Cisco DCNM to provision and monitor the Cisco Nexus 5000 and 7000 Series Switches (Data Center Network Manager). Use Cisco Fabric Manager to manage the Cisco MDS 9000 Family of Fibre Channel switches (Fabric Manager Web Client).

Figure 15 Data Center Network Manager

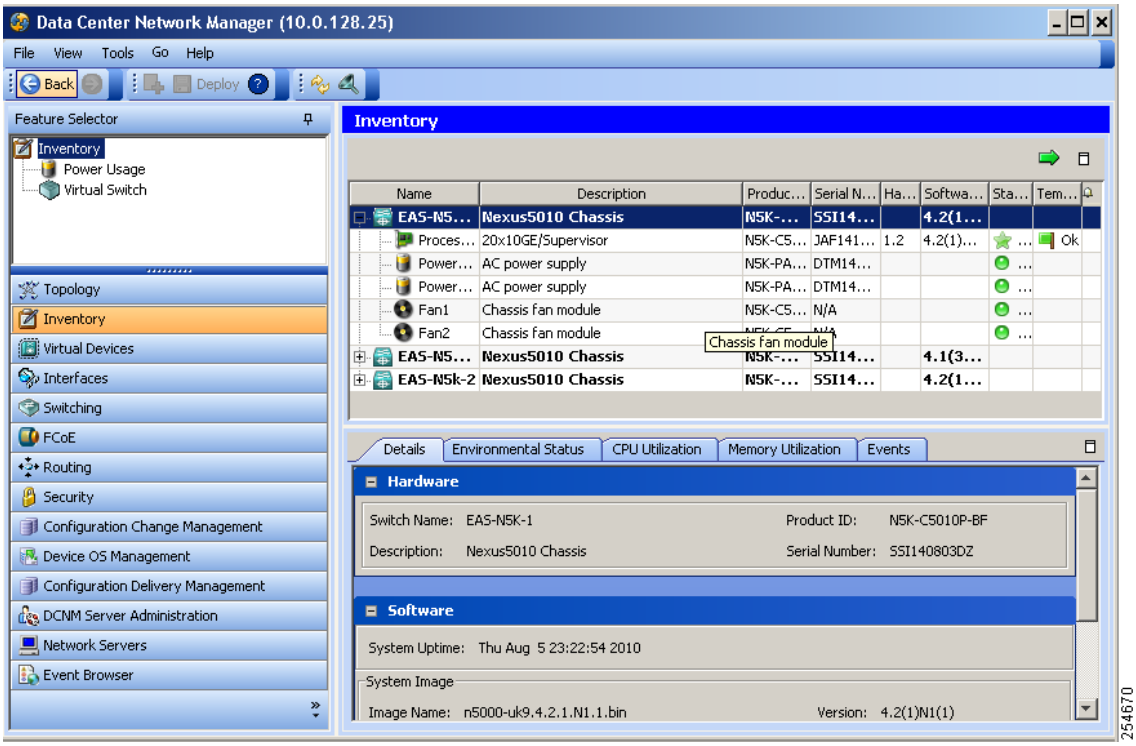
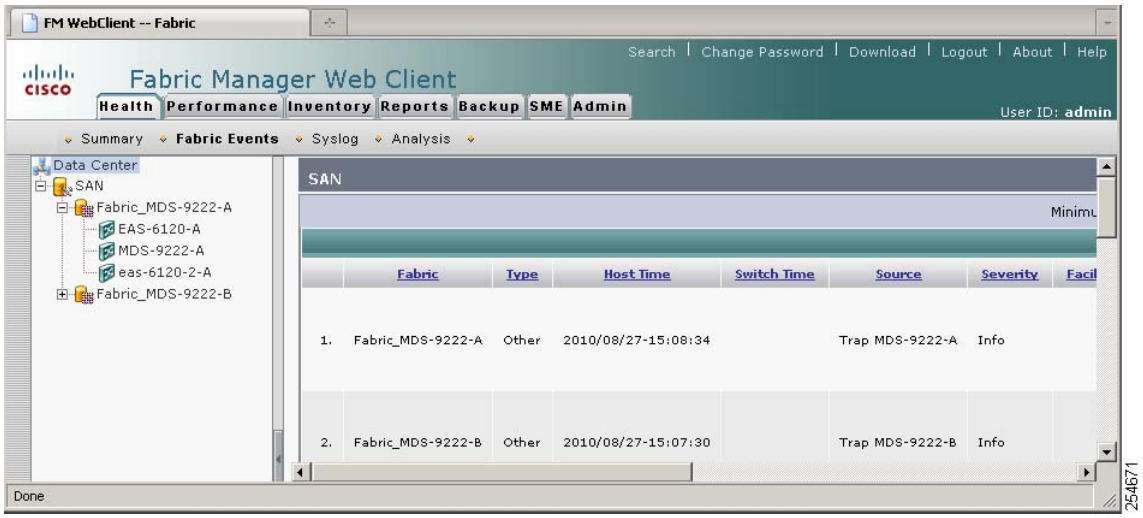
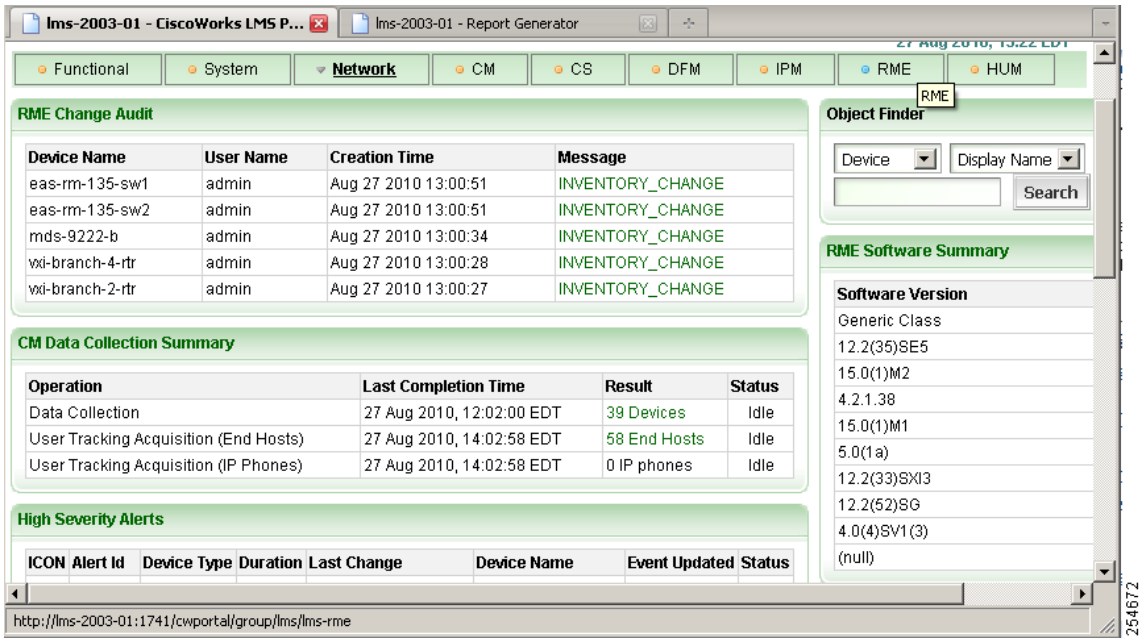


Figure 16 Fabric Manager Web Client



Use CiscoWorks LMS to monitor and provision routers, switches, and other network devices throughout the end-to-end system (Figure 17). It can be used to measure network link utilization, latency, and jitter on all managed interfaces.

Figure 17 CiscoWorks LMS Resource Manager Essentials



The Cisco Unified Communications Management Suite is the recommended tool for managing Cisco Unified Communications Manager, Cisco Unified Presence, Cisco Unity Servers, Cisco Unified Personal Communicator client, Cisco Cius Tablet, and Cisco IP Phones deployed in the system. In addition to the embedded web-based management interfaces available with the Cisco Unified Communications suite of servers, the Cisco Unified Communications Management Suite is recommended for large-scale deployments and consists of the following tools: Cisco Unified Operations Manager, Provisioning Manager, Service Monitor, and Statistics Manager.