Cisco Virtual Workspace (VXI) Smart Solution 2.7 Performance and Capacity Results Guide for VMware

May 3, 2013

Contents

Contents 1

Introduction 2

Compute and Storage 2

- Single Server Scale and Performance Results 5
  - Summary of Results 5
  - Validation Methodology 6
  - Detailed Test Results 8
  - HVD Scalability on Cisco UCS B200 M3 with VMware View 5.1 8
  - VMware View Storage Accelerator 12
  - Vblock Profile 15
  - FlexPod Profile 21
  - View4.6/ESXi4.1/RDP/B250M2 Profile –New CPU Counter 26
  - View4.6/ESXi4.1/PCoIP/B250M2 Profile –New CPU Counter 31
  - View5/ESXi5/RDP/B250M2 Profile - Vblock 36
  - View5/ESXi5/PCoIP/B250M2 Profile - Vblock 40
  - HVD Scalability for View5/ESXi5.0 Profile on UCS B230 M2 46
  - Scale and Performance Baseline for VMware View without Antivirus 49

Network Characterization 54

- Summary of Results 54

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

Validation Methodology 55
Detailed Test Results 55
Bandwidth Characteristics of a DV workload – Cisco KW+ workload 56
Bandwidth Characteristics of a Video Only DV workload 62
Impact of Protocol Adaptiveness on Server/Compute Performance 67
Impact of WAAS Optimization on Cisco Virtual Workspace WAN deployments with View RDP 70
Rich Media Application Characterization 74
Summary of Results 74
Validation Methodology 75
Detailed Test Results 75
Detailed Performance Results 87

Introduction

This performance guide serves as an addendum to the Cisco Virtual Workspace (VXI) Smart Solution 2.7 with VMware View 5.1 located here:

http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VXI/CVD/VXI_CVD_VMware.html

The primary objective of this guide is to provide a detailed analysis of the results from the various scale, performance, and other characterization testing done in the end-to-end Cisco Virtual Workspace system. The results in this document can provide key data points that can be used in your environment for capacity planning, particularly for estimating the sizing of various components that make up a Cisco Virtual Workspace system. However, the results presented here are based on a given workload that may not be representative of the workload generated by your user base. Readers are therefore advised to carefully consider their own workloads and make adjustments to the estimations as needed to suit the needs of their deployment.

This document is organized into three main sections each focused on providing capacity planning data relevant to key subsystems in the larger Cisco Virtual Workspace system, namely Compute and Storage, Network and Applications that provide Rich Media experience.

Compute and Storage

In this section we look at the sizing data for the compute and storage aspects of the Cisco Virtual Workspace system based on the testing done in the end-to-end system. The primary focus is on characterizing the scalability and performance Cisco UCS servers (B-series and C-series) for different deployment profiles commonly seen in virtualized desktop environments. The specific models of UCS servers characterized in this guide are UCS B200 M3, UCS B230 M2, UCS B250 M2 and UCS C250 M2 though several other models of the UCS B-series and C-series are supported in the solution. Processors deployed for these servers are typically the best processor model available during the time of validation. Details of the server and the model used for each test profile are provided with the results in the next section. Memory configurations used is based on the configurations recommended in the Cisco Virtual Workspace Offer bundles which are typically 256 GB to 384 GB of memory.

The hypervisor used for all deployment profiles is ESXi with sizing and performance data available for ESXi 4.1, ESXi 5.0 and ESXi 5.0U1. ESXi 5.0 provides a number of optimizations that can greatly improve the scalability of any ESXi based deployment of hosted virtual desktops. One such feature is the adjustment of the HaltingIdleMsecPenalty (HIMP) parameter which affects the algorithm that grants access to CPU resources. In vSphere 5.0, this kernel adjustment is enabled by default and improves the fairness for virtual desktops particularly under load. To quantify the impact of these optimizations on Cisco UCS servers, testing was done in the Cisco Virtual Workspace system to determine the density improvement. Results from this testing are included in the Single Server Scale and Performance section of this document. See VMware KB article http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1020233 for more details on this optimization.

ESXi has a number of advanced memory reclamation and management capabilities that enable the host physical memory to be over-committed. These include features such as transparent page sharing, memory ballooning, memory compression, and hypervisor swapping. As you go through the capacity planning process, it is important to review the memory related chapters in vSphere Resource Management Guide that is published with every release of vSphere, to understand how these features take effect particularly as you over-commit memory. One such feature that Cisco Virtual Workspace leverages is Transparent Page Sharing which comes into play, typically only at densities higher than 100. However the over-commitment is usually below 5% due to the workload (see below) and physical memory used in validation. Other features such as memory ballooning and swapping to disk by the hypervisor are monitored during testing but in this case, it is done to ensure they are not in effect per the success criteria used for scale and performance testing in Cisco VXI. Nevertheless, these are fail safe mechanisms built into ESXi that come into play as memory gets more and more over-committed to prevent complete server failure that can impact all virtual desktop users on that server.

ESXi also reserves 6% memory for hypervisor use but this can be reduced to 2% in servers that have more than 64G of memory. Cisco Virtual Workspace did not leverage this feature but it can be enabled on all Cisco UCS servers. See VMware KB article http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1033687 for more details on this option.

One change worth noting about the performance data in this document is that the counter used for monitoring the CPU utilization has been changed in recent testing based on recommendations from VMware. CPU Utilization of 90% is one of the success criteria used for determining the number of virtual desktops a given UCS server can support and changing this counter has improved the VM density that the server can support. Application response times used as a gauge of user experience (UE) were still within the acceptable range – see next section for details on this specific success criterion. Performance data based on the older and newer counter are both included in this document and it is important to keep this change in mind as you go through the results. Cisco Virtual Workspace also did targeted testing to quantify the impact of this counter change. These results are included in the Detailed Results section below.

Note All performance data with ESXi5 was based on the newer CPU utilization counter, however, utilization data from both counters are included in the CPU charts provided with these results.

Virtual Desktops are delivered using VMware View and Linked clones in the Cisco Virtual Workspace system and data for both View 4.6, View 5.0, and View 5.1 are included in this document. Linked clones are highly recommended for the storage capacity savings it provides since a major hindrance to desktop virtualization adoptions is the high cost of storage. The storage savings come from the individual desktops sharing a common base image called a Replica (or Parent). Each desktop uses a desktop image that comprises of the larger shared common base image and a small differential image that uniquely defines the desktop (clone) such as the hostname, IP etc. As the desktop is used, any deltas from the parent image are captured in each clone's differential disk. The differential disks are also thin

provisioned which maximizes storage use by allocating space only as needed. Using linked clones can therefore drastically reduce the storage needs from that of a full desktop. A deployment where the desktop's operating system takes up 10-15G of disk space no longer needs this amount of space on a per desktop basis if it can share that portion of the image from the parent OS image. So a deployment of 100 desktop that would've needed 100x10G=1TB of space will now need 1x10G of space for the OS image with a differential disk that uses a fraction of the disk space that a full desktop would've needed.

A key trait of this architecture is the separation of the main OS disk (Read-only) from delta disk that captures all changes a user makes. This results in a read intensive workload on the storage system when users bootup or login to their desktops by accessing the same parent image. However, View provides Storage tiering, where the virtual desktop composed of the replica and linked clone can then be placed on low capacity but high performing SSD drives while the clones as well as the user data can be placed on less expensive, high capacity storage such as SAS or SATA drives. Use of SSD drives to house the replica serving a large pool of linked clone desktops will improve the user's desktop experience significantly. Cisco Virtual Workspace system recommends and uses this architecture for validation in cases where the storage vendor supports this.

Linked clones can also be setup to be persistent where the state of the changes to desktop is retained or non-persistent where the desktop is refreshed back to the original state at log-off. The Cisco Virtual Workspace system validation is primarily targeted at a Knowledge Worker type user and as such persistent desktops are deployed. The workload used for all the testing documented here is using a Cisco Knowledge Worker+ (KW+) workload. This includes not only standard applications such as Microsoft Office, Adobe and Internet Explorer but also includes a Cisco rich media application and a hypervisor based antivirus solution as a part of the Cisco KW+ workload. A detailed overview of this workload can be found under Workload considerations in the Performance and Capacity chapter of the Cisco Virtual Workspace (VXI) Smart Solution 2.7 with VMware View 5.1. Note that the version of the KW+ workload script used for each test profile is included in the Summary of Results table in the next section since it would be important to know if a different workload was used particularly when comparing results.

With persistent desktops, delta disks of the clones can grow in size and become as big as the parent disk. A well managed environment can refresh the OS disk back to parent image to keep this from happening and provide persistency for any changes the user makes to the desktop through other means, namely user profiles. User profile portion of the desktop stack can also be decoupled from the virtual desktop with user profile virtualization such that the user is assigned a generic desktop at login but with this capability, the desktop that the user logs on is no different from a desktop dedicated to the user.

Cisco Virtual Workspace system supports both shared storage (NAS, SAN) and Direct Attached Storage (DAS). For shared storage, the storage array used in the Cisco Virtual Workspace system is either an VSPEX (EMC VNX 5500) or a NetApp (FAS 3170). For validation, EMC is deployed as a Fiber Channel attached SAN while NetApp is used as either NFS or iSCSI based storage though other storage connectivity options are available and supported in the Cisco Virtual Workspace system. The storage arrays are deployed in a highly scalable storage architecture based on best practices and recommendations from EMC and NetApp.

For DAS storage, local disks on the UCS servers are used and they can be SATA, SAS, SSD or a combination of these. DAS for virtual desktop deployments is a lower cost option but should be used with a careful consideration of the use case and the features that you loose as a result such as high availability, load balancing and vMotion. Another consideration is whether the target deployment needs a persistent or a non-persistent desktop as local disks are fairly limited in size and can be use for the storing base desktop but typically not for per-user customizations, user-installed applications or user data.

A key feature worth mentioning due to its significance to virtual desktop workloads is the use of a tiered caching or storage as a part of your storage architecture. Using RAM based or SSD based caching can significantly benefit desktop virtualization (DV) workloads as the parent image will likely get served by

the cache after the first desktop boots up. This will minimize the impact of login storms or boot storms where the Read IOPS tend to be high as the IOPS will be served by the cache rather than by the backend disks. This will help reduce the number of disks required to meet Read IO performance particularly during Bootup and Login of a large pool of desktops.

Storage and Performance Optimization solution from Atlantis ILIO can also optimize both the Read and Write IO traffic from the desktop and significantly reduce the IO load on the back-end disks. It can also reduce the overall storage capacity needs by optimizing the Write IO traffic in addition to the Read IO. Results from the testing done in the Cisco Virtual Workspace system is also included below.

Please refer to the Performance and Capacity chapter of the Cisco Virtual Workspace (VXI) Smart Solution 2.7 with VMware View 5.1 for a more comprehensive overview of the planning process, design considerations, and best practices.

Single Server Scale and Performance Results

This section covers the following aspects of the scale and performance testing done in the Cisco Virtual Workspace system:

- High level summary of deployment profiles tested
- Validation methodology
- Detailed test results

Summary of Results

In this section, a high level summary of the deployment profiles characterized from a single server scale (SSS) perspective across the end-to-end Cisco Virtual Workspace system are provided in the table below. The primary objective of each test is also provided in the rows preceding the profile information.

Table 1 Profile Information

Objective	Server Model	Storage	Desktop Virtualization Profile	HVD Profile
Scalability and performance characterization of Cisco UCS B200M3 server with VMware View (Vblock)	Cisco UCS B200 M3 with 384G of memory	VSPEX (EMC VNX 5500) - Fibre Channel	VMware View 5.1 on VMware ESXi5.0U1	Microsoft Windows 7 32-bit with 2 GB of memory and 20 GB disk; Persistent
Scalability and performance characterization of Cisco UCS B230M2 server (Vblock)	Cisco UCS B-230 M2 with 256G of memory	VSPEX (EMC VNX 5500) - Fibre Channel	VMware View 5.0 on VMware ESXi 5.0	Microsoft Windows 7 32b with 1.5G of memory and 20G disk; Persistent
Scalability and performance characterization of Cisco UCS B250M2 server (Vblock)	Cisco UCS B-250 M2 with 192G of memory	VSPEX (EMC VNX 5500) - Fibre Channel	VMware View 4.6 on VMware ESXi 4.1	Microsoft Windows 7 32b with 1.5G of memory and 20G disk; Persistent

Objective	Server Model	Storage	Desktop Virtualization Profile	HVD Profile
Scalability and performance characterization on Cisco UCS B250 M2- Impact of success criteria (CPU utilization counter changed) and vSphere 5.0 changes (Vblock)	Cisco UCS B-250 M2 with 192G of memory	VSPEX (EMC VNX 5500) - Fibre Channel	VMware View 5.0 on VMware ESXi 5.0 RDP & PCoIP	Microsoft Windows 7 32b with 1.5G of memory and 20G disk; Persistent
Storage Optimization with VMware's View Storage Accelerator	Cisco UCS B200 M3 with 384G of memory	VSPEX (EMC VNX 5500) - Fibre Channel	VMware View 5.1 on VMware ESXi5.0U1	Microsoft Windows 7 32b with 1.5G of memory and 20G disk; Persistent

KWP is the internal designation given to the automated workload used to simulate a user's activities on a desktop.

Note Results published are specific to the Cisco Virtual Workspace architecture including Cisco UCS class of blade servers. Per chassis density ratio supported by RDP or PCoIP protocols can vary based on the overall desktop virtualization architecture.

Validation Methodology

In this section we take a look at the validation methodology used in the scale and performance testing done in the Cisco Virtual Workspace system. All of the above testing was done across an end-to-end Cisco Virtual Workspace network based on the Cisco Virtual Workspace system architecture outlined in the Cisco Virtual Workspace (VXI) Smart Solution 2.7 with VMware View 5.1 document.

Workload Profile: Cisco Knowledge Worker+

The workload used is a critical factor for any performance related characterization done in a desktop virtualization environment. All the test results presented in this document were done using the Cisco KnowledgeWorker (KW+) workload unless stated otherwise. An overview of this profile is provided in the Workload considerations section of the Performance and Capacity Planning chapter in the Cisco Virtual Workspace (VXI) Smart Solution 2.7 with VMware View 5.1 document. Cisco KW+ workload also includes a hypervisor based optimized antivirus solution from a leading vendor.

All testing was done using Test and Performance Platform (TPP) from Scapa Technologies. This tool is used for all scale, performance and other characterization type testing to initiate a large number of user sessions and execute a workload across these sessions.

Note The test tool used for a given test is not particularly important as long as the workload it implements it is representative of the type of users it is designed to emulate. As such, Scapa is implementing a workload representative of a Knowledge Worker but in addition to that, it also includes antivirus and a Rich Media application in the workload and hence the term KW+. A close evaluation of the workload profile (see above) and the results will show that this is in fact the case.

Success Criteria

The success criteria can vary depending on the specific objective of the test. But for the most part, if the objective is to determine the virtual desktop density that can be supported on a given model of the server for the specified deployment profile using a Cisco KW+ workload profile, then the success criteria typically used are as follows:

- Good User Experience based on application response times – see next section
- CPU Utilization of 80% and/or 90%
- Memory Utilization of 90% with no ballooning (ESXi) or host swapping with some exceptions
- Average IO Latency less than 20 ms

Application Response Times

The table below summarizes the average application response times used as the success criteria for the performance testing done in the Cisco Virtual Workspace system. On each virtual desktop hosted on the UCS server, Scapa load generation tool will initiate a VDI session and then initiate activities defined in the workload profile to generate a workload on each desktop. Applications in the workload (except for Cisco Unified Personal Communicator in deskphone mode) are launched and closed in every iteration of the workload loop. Therefore the average response times measured (shown below) for a given application is a combination of the response times measured for that application across all HVDs running on a server as well as the response times across multiple iterations of the workload running on each HVD. The success criterion was derived from a combination of testing done on physical desktops and HVD with these applications and measuring the response times. For each test, the response times measured are compared against the success criteria defined below in order for the test to pass. It is also important to note that Scapa measures the response times from a user/endpoint perspective and not from the hosted virtual desktop in the data center when the display protocol is RDP or ICA. For PCoIP, it is measured at the virtual desktop in the data center – this is typical of most load general tools.

Table 2 Success Criteria.

Applications	Success Criteria for Maximum Acceptable Startup Times
Cisco Unified Personal Communicator 8.5 in deskphone control mode	5s
Outlook	5s*
Excel	5s
PowerPoint	5s
Acrobat	5s
Internet Explorer	5s
Word	5s*
* Testing in previous releases of Cisco Virtual Workspace used a 10s success criteria	

Performance Metrics

The following aspects of the server performance are measured for each deployment profile tested. For ESXi, esxtop is used to measure these metrics using a 5s polling interval. Storage statistics from NetApp and EMC are included where possible.

- Average CPU Utilization

- Average Memory Utilization
- Storage
  - IOPS
  - IO Bandwidth
  - IO Latency
- Network Bandwidth Utilization

Detailed Test Results

A detailed analysis of the test results and the associated profile and objectives are provided in this section.

HVD Scalability on Cisco UCS B200 M3 with VMware View 5.1

When deploying a Cisco Virtual Workspace (VXI) Smart Solution based on VMware View and Cisco UCS servers, it is critical to understand the scalability and performance of the physical server hosting the desktops. The server scalability in terms of the number of desktops supported on a single server will determine the total number of servers needed for the deployment. The storage (I/O, I/O bandwidth, I/O latency) and network bandwidth metric measured from a fully loaded server can be used to size the storage and data center network links for the overall deployment.

The results provided in this section are based on the testing done on a Cisco UCS B200 M3 server in an end-to-end Cisco Virtual Workspace Smart Solution using Vblock infrastructure running VMware View 5.1, ESXi 5.1 and VSPEX (EMC VNX 5500). Results indicate that ~125 Microsoft Windows 7 32-bit virtual desktops can be supported on a Cisco UCS B200M3 using Cisco Knowledge Worker + (KW+) workload. Response times for most applications in the workload is <2sec with one application having a response time in the 2-3s range.

Results also indicate that we are CPU bound for this profile with 384GB (2GB per desktop) of memory deployed per server. A Cisco UCS B200M3 can support up to 768GB of memory with 32GB DIMMS and 384GB of memory with 16 GB DIMMS. When a Cisco UCS B200 M3 server is deployed for user desktops, Cisco generally recommends a performance optimized memory configuration of 256GB, particularly with a 1.5GB per desktop allocation. The same could've been done for this testing by allowing for memory over-subscription. However, the results here provide data based on CPU limit for customers that may choose to size their deployment by adding more memory to their servers rather than the two alternative options of [1] same density but with memory over-subscription at 256GB of memory or [2] lower density without memory over-subscription at 256GB.

Detailed Performance Results

This section provides a detailed overview of the test setup and results in terms of the configuration, performance charts, and application response times for supporting 125 desktops on a Cisco UCS B200 M3 server.

Summary of Test Results

Using the above deployment profile, 125 VMs can be supported on a Cisco UCS B200M3 with the following performance metrics.

- Average CPU Utilization = ~90% (Steady state)
- Average Memory Utilization based on allocated memory = ~70%

- Average I/O Latency <2ms
- Application Response times <3sec

Test Profile

This section provides configuration, environment and setup details used in this testing.

Desktop Virtualization

- VMware View 5 .1
- Connection protocol – PCoIP
- Linked clones

Hypervisor

VMware ESXi 5.1

Virtual Desktop Configuration

- Microsoft Windows 7 32-bit desktops with 2 GB of RAM, 20 GB disk and 1vCPU per desktop
- Persistent desktops

Server Specifications

- Cisco UCS B200 M3 with Dual 8-core Intel Xeon E5-2690 processors @ 2.90 GHz and 384GB RAM (24 x 16GB DIMMS @ 1666MHz)
- Cisco UCS VIC 1240 Virtual Interface Card- 4x10Gb

Workload Profile: Cisco Knowledge Worker+ (ver4.25)

- Microsoft Office 2010 Applications
- Internet Explorer
- Adobe Acrobat 9
- Cisco Jabber for Windows – Version 9.1.3
- Optimized antivirus solution from a leading vendor
- 30 second Flash Video

Storage

VSPEX (EMC VNX 5500) – Fibre Channel

Data Collection/Test Tool

- Workload Generation Tool - Scapa Test Performance Platform (TPP)
- Resxtop with a polling interval of 5s
- End user response times measured using Scapa
- Data collected for Login, Workload and Logout stages

Performance Charts

Application Response Times

The table below shows that the response time experienced by 125 users were well within 5sec or the established success criteria for all applications.

Table 3 Response Times for 125 desktops on Cisco UCS B200M3 with View5.1/ESX5.1/PCoIP/EMC VNX

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 125 desktops on Cisco UCS B200M3
Cisco Jabber for Windows (Version 9.1.3)	5s	1.6s
Outlook'10	5s	2.7s
Word'10	5s	1.8s
Excel'10	5s	1.6s
PowerPoint'10	5s	1.7s
Internet Explorer	5s	1.6s
Acrobat	5s	1.6s

Server Performance

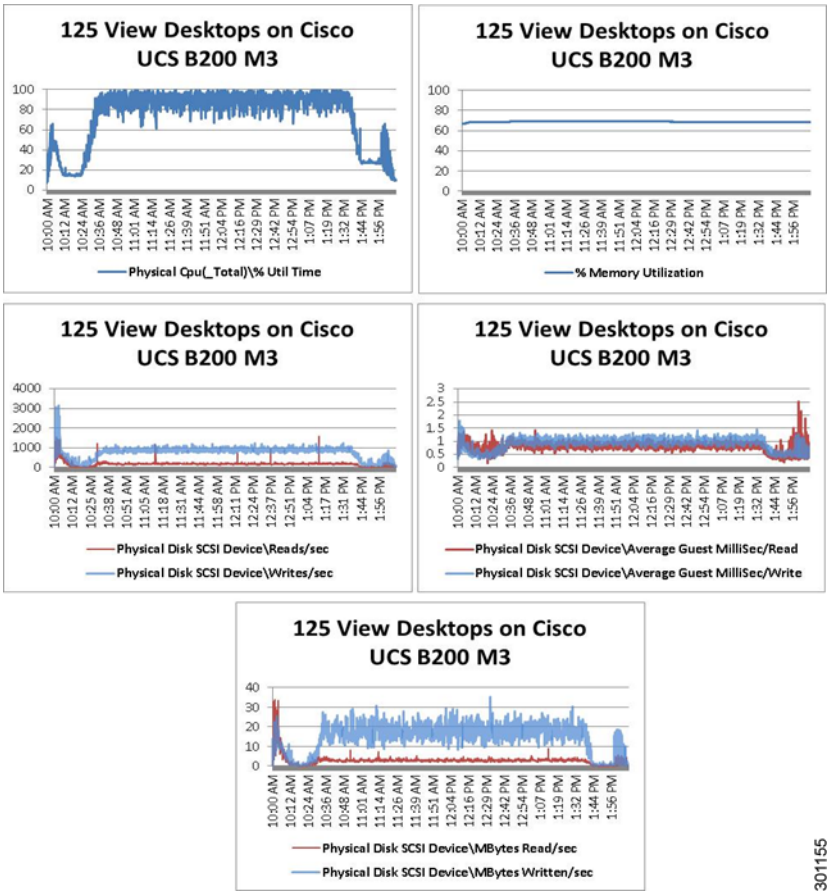
The overall performance of a Cisco UCS B200M3 server in terms of the CPU utilization, memory utilization, I/O load, I/O performance and Bandwidth generated by 125 View desktops running a Cisco KW+ workload are shown in the figure below.

The first chart shows the CPU utilization measured using resxtop with a 5s polling interval during the Login, Workload and Logout stages with 125 desktops on a Cisco UCS B200 M3 server. This chart confirms that we are CPU bound for this profile with a CPU utilization of ~ 90% during steady state use of their desktop by 125 users.

The second chart is the memory utilization chart showing the memory allocated to the hypervisor and virtual desktops which is at ~70%. This represents the actual memory allocated to the desktops based on resxtop data and not what was actually used by the desktops. Note that 384G of memory was deployed on the server with a 2GB per desktop allocation as 2GB of memory is recommended for Cisco Jabber application running on the desktop.

The next few charts show the storage performance, in terms of read and write I/O load on the storage system generated by a single server of users running the workload. The read and write I/O load profile is typical of a VDI workload. The I/O latency is <3msec for both peak and average – we typically aim for an average I/O latency < 20ms. The I/O Bandwidth data chart shows the network bandwidth utilization associated with the storage traffic – note that this is in Mbytes/sec and can be a starting point for estimating the bandwidth needs in the data center.

Figure 1 Server Performance for 125 desktops on Cisco UCS B200M3 with View5.1/ESX5.0U1/PCoIP/EMC VNX



The peak and average I/O performance data shown in the charts above are also summarized in the table below.

Table 4 I/O Performance during Steady State, Login and Logout of 125 users

Storage I/O	Steady State	Login	Logout
Read-Avg	188.91	535.74	39.18
Read-Peak	1568.96	1549.49	260.56
Write-Avg	890.61	884.82	223.23
Write-Peak	1256.16	3145.32	860.01
Read-Latency-Avg.	0.83	0.91	0.56
Read-Latency-Peak	1.42	1.37	2.52
Write-Latency-Avg.	1.02	0.78	0.55
Write-Latency-Peak	1.44	1.79	1.28

The I/O data shows the steady state read and write IOPS are approximately 900 write IOPS and 200 read IOPS for 130 desktops – the ratio seen here is pretty close to what is expected during the workload stage of a VDI workload. A virtual desktop workload during Steady State is typically 80% writes and 20% reads but can vary by 5-10% in either direction as seen in this testing. This workload used in this test

generates approximately 9 IOPS per desktop. Typically knowledge worker workloads may generate higher IOPS per desktop depending on the environment, applications deployed and optimizations in place. From a deployment planning perspective, Enterprises should assess their environment and users for a more accurate estimation of the IOPS per desktop needed for their deployment.

In summary, the single server scalability of a Cisco UCS B200M3 with VMware View 5.1 and ESXi5.1 and running Cisco KW+ workload is 125 desktops with CPU being the limiting factor. Memory is not a limiting factor as the Cisco UCS B200 M3 can support up to 768 GB of memory using 32 GB DIMMs.

VMware View Storage Accelerator

Cisco Virtual Workspace delivers significant value to customers by integrating solutions and technologies that reduce Total Cost of Ownership (TCO) and improve Return on Investment (ROI) for virtual desktop deployments. Storage costs are a significant contributor to the overall cost of a virtual desktop and VMware's View Storage Accelerator is a solution that can substantially reduce the storage costs by reducing the performance requirements on the backend storage array.

VMware's View Storage Accelerator (VSA) is a storage performance optimization solution that leverages a hypervisor or a host based cache to cache the data previously accessed from the storage array, by the virtual desktops on that host. VSA is based on a vSphere 5.0 caching feature called Content Based Read Cache (CBRC). As the name suggests, VSA is designed to offload the Read requests going to the storage array by serving the data (once cached) from the local read cache maintained by the hypervisor. VSA is targeted for View deployments and enables View desktops to leverage the hypervisor cache. VSA is available as of View 5.1 and uses the server (Cisco UCS) memory as its read cache.

VSA validation in Cisco Virtual Workspace focused on understanding the storage performance improvements that can be achieved with 160 View desktops deployed on a single Cisco UCS B200M3 server and using default cache settings. VDI workloads are typically Read I/O intensive during Boot up, Login and Application launches so testing was specifically done for these stages. 160 View desktops were booted up and logged in for the VSA testing. The login stage per Cisco KW+ workload also includes a pre-run of all applications in the workload and simulates and therefore the login of 160 desktops for the test also includes an application launch of all applications in workload across 160 desktops. Note that 160 desktops were deployed for this test because it represents the max scalability achieved on a Cisco UCS B200M3 based on the testing done in the Cisco Virtual Workspace system using Cisco KW+ workload.

Results Analysis Summary

- Using VMware's VSA feature, results from the testing shows that it significantly offloads the read I/O going to the storage array. Testing indicates a Read I/O offload of over 80% during login of 160 desktops on a Cisco UCS B200 M3 server. A 65% reduction in average Read I/O and 40% reduction is peak Read I/O was also seen during the boot up of 160 desktops. Higher levels may be possible with different boot and login profiles, larger cache sizes or by enabling VSA on persistent disks. VSA is enabled by default only on OS disks and this was the case for this testing as well. It is important to note that for a larger deployment with many servers, enabling this caching on all servers will have a far more significant impact in reducing the overall performance needed from the storage array, thereby reducing storage costs.

Table 5 Read IOPS Offload with VMware VSA

160 View Desktops on a UCS B200M3	Without VSA (To Storage array)	With VSA (To Storage array)	IOPS Offload by VSA
Peak Read IOPS (Boot up of 160 desktops)	~28,000	~17,000	~40%
Average Read IOPS (Boot up of 160 desktops)	~8360	~2780	~67%
Peak Read IOPS (Login of 160 desktops)	~9740	~1675	~83%
Average Read IOPS (Login of 160 desktops)	~4100	~730	~82%

- Testing also showed that boot up and login times are reduced by minutes during the login and boot up of 160 desktops. Results indicate that boot up was complete in ~2min using VSA where as it took ~5min without VSA. Similarly, login time went down to ~10min with VSA from ~13 min without VSA. Though this reduction is across all 160 desktops, it should still have a positive impact from a user experience perspective, particularly during boot up and login, but also during desktop use.
- Another benefit worth mentioning is that VSA will reduce the network bandwidth needed for storage traffic. This is inherent in the feature since the requests are served from on-board cache, resulting in less I/O traversing the network to and from the storage array.

Design and Deployment Considerations

Note Note that host caching is not in effect or used by virtual desktops until it is enabled in View though the host may be enabled for it from vCenter.

- Feature can be enabled during pool creation or by editing pool settings from View Administrator interface but desktops require a reboot for the changes to take effect.
- View supports both OS disks and persistent disks – both can be enabled for host caching; default is OS Disk only
- RAM based cache – does not survive reboot
  - Rebuilding of cache starts as the first virtual desktop boots up
  - Boot up of first desktop should cause the master image or replica to be loaded into cache
  - Virtual desktops from the same pool and on the same host can now use the replica from cache resulting in faster boot ups
- For continuous benefit, periodic cache regeneration is recommended to flush invalid data caused by write actions to a desktop's VMDK file. Due to CPU impact, regeneration should be done during non-production or periods of least use.
- Benefit may reduce past login phase since virtual desktop workloads are typically write I/O intensive post-login. However the benefit of minimizing peaks in Read I/O is significant when sizing storage as peaks during boot up of 160 desktops for a 5min period can be over 25,000 Read IOPS with the boot profile used for this testing. See Performance charts below for more details.

Detailed Performance Results

This section provides a detailed overview of the test setup and results in terms of the configuration, performance charts, and improved user experience with VSA enabled for 160 desktops on a Cisco UCS B200 M3 server.

Test Profile

This section provides configuration, environment and setup details used in this testing.

Desktop Virtualization

VMware View 5 .1 Linked Clones using PCoIP

Hypervisor

VMware ESXi 5.0 U1

Virtual Desktop Configuration

- Windows 7 32b desktops with 1.5G of RAM, 20G disk and 1vCPU per desktop
- Persistent desktops

Server Specifications

- Cisco UCS B200 M3 Server with Dual Eight Core Intel Xeon E5-2690 processors @ 2.90 GHz and 384G RAM (24 x 16GB DIMMS @ 1666MHz)
- UCS VIC 1240 Virtual Interface Card- 4x10Gb

Workload Profile: Cisco Knowledge Worker+ (ver3.3)

- Microsoft Office 2010 Applications
- Internet Explorer
- Adobe Acrobat 9
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

VSPEX (EMC VNX 5500) - Fibre Channel

View VSA Configuration

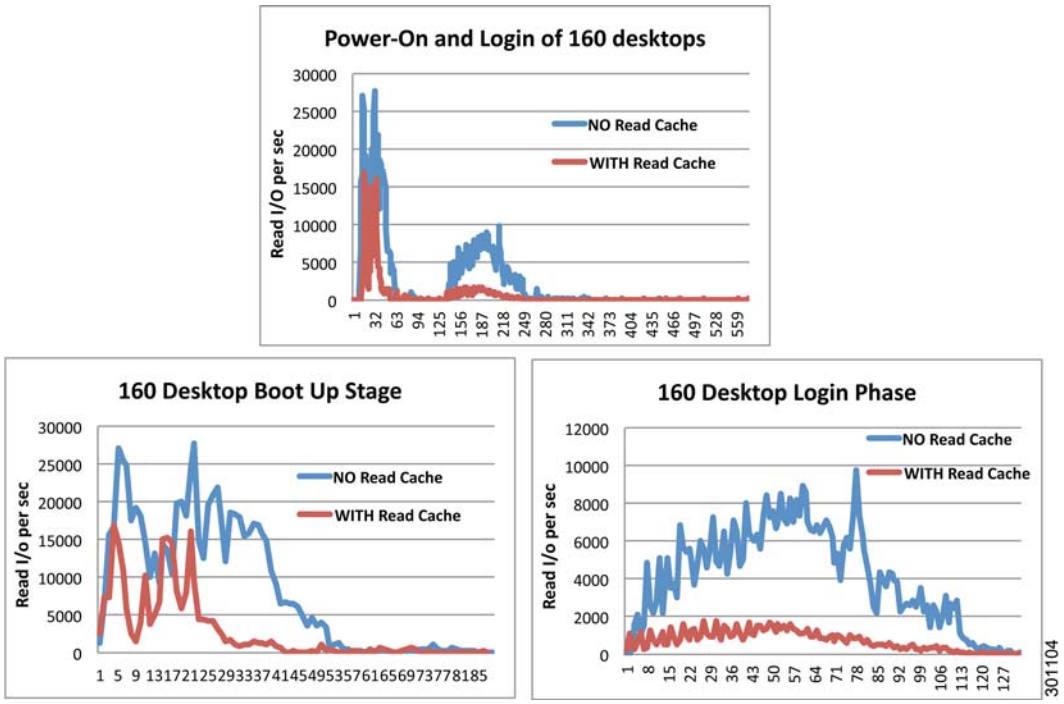
- Default cache size of 1MB was used
- Caching was enabled only on OS disk

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s
- End user response times measured using Scapa TPP as outlined earlier in this document
- Data is captured and graphed for Login and Workload stages

Performance Charts

Figure 2 Read I/O and User Experience improvements with VSA for 160 View desktops on Cisco UCS B200M3



The top graph shows the boot up and login of 160 View desktops on a Cisco UCS server with and without VSA enabled. This graph is important in that it shows, at a high level, the relative read I/O load during boot up and login. For the setup used in this testing, as stated earlier, the login Read I/O also includes launching of all applications in the workload. This is done in Cisco Virtual Workspace testing to setup the application environment prior to entering the workload stage. However, in this particular test, it is useful to quantify the I/O benefit with application launches since they occur often as users use their desktops. Here these events are occurring in the Login stage but the benefit is still important to understand.

The bottom two charts show a detailed view of the Boot up and Login stages. These graphs clearly show the reduction in I/O going to the back end shared storage, but also show the impact to login and boot up times. Each sample on the x-axis represents a 5s progression in time based on the polling interval being 5s for these tests. Based on this, using the charton the far right corner, it can be clearly seen that the reduction in login time for logging in 160 desktops is approximately 2min (((115-90)*5s)/60s) by using this feature.

In conclusion, due to the significant reduction in Read I/O and improvement in user experience, VSA should be leveraged when possible for a scalable View deployment with reduced storage costs.

Vblock Profile

This section covers the results of the single server scalability tests done for a UCS B250 M2 using a Vblock infrastructure environment with Windows 7 32b desktops running on View 4.5 and ESXi. Based on the testing done with Cisco KW+ workload, 85 virtual desktops (@ 85% CPU) can be supported on a UCS B250 M2 using this deployment profile. Higher VM density is possible at 90% utilization.

Test Environment and Setup

- View 4.5 on ESXi 4.1; RDP
- HVD Profile:
  - Windows 7 32b with 1.5G of memory and 20G of disk space
  - 1 vCPU, Persistent desktop
- Workload Profile: Cisco KW+ (Cisco Unified Personal Communicator 8.5 in deskphone mode, IE, Microsoft Office 2007 Apps, Acrobat) with optimized antivirus solution from a leading vendor
- UCS Server: B250 M2 with 192 G of memory - Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- Storage: SAN based using VSPEX (EMC VNX 5500)
- All of the data shown in the graph below is collected using resxtop with a polling interval of 5s except for IO Statistics from EMC
- All response times are measured using Scapa TPP as outlined above
- For this profile, data is captured and graphed only for the workload phase once it reached steady state

Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this section.

Table 6 Application Response Times for View 4.5 on ESXi on a UCS B250M2 with EMC - Vblock

Applications	Success Criteria for Maximum Acceptable Startup Times	Average Startup Times Measured (sec)
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	1.8s
Outlook	10s	2.3s
Excel	5s	.9s
PowerPoint	5s	.7s
Acrobat	5s	.4s
Internet Explorer	5s	3.3s
Word	10s	7.2s

Summary of Test Results

For the deployment profile detailed in the Test Environment and Setup section above, 85VMs can be supported on a Cisco UCS B250 M2 with the following performance metrics. For this test run, VM density was also measured at ~85% CPU utilization while benchmarking tests typically use 90%. A higher number of virtual desktops can be supported with this profile at 90% CPU.

- Average CPU Utilization = 85% (Steady state)
- Average Memory Utilization = 70%

- Storage – Peak values not captured as data was captured once workload reached steady state
- Network Bandwidth Utilization = < 20Mbps during steady state

Figure 3 CPU Utilization for View 4.5 (RDP) on a UCS B250 M2 using EMC

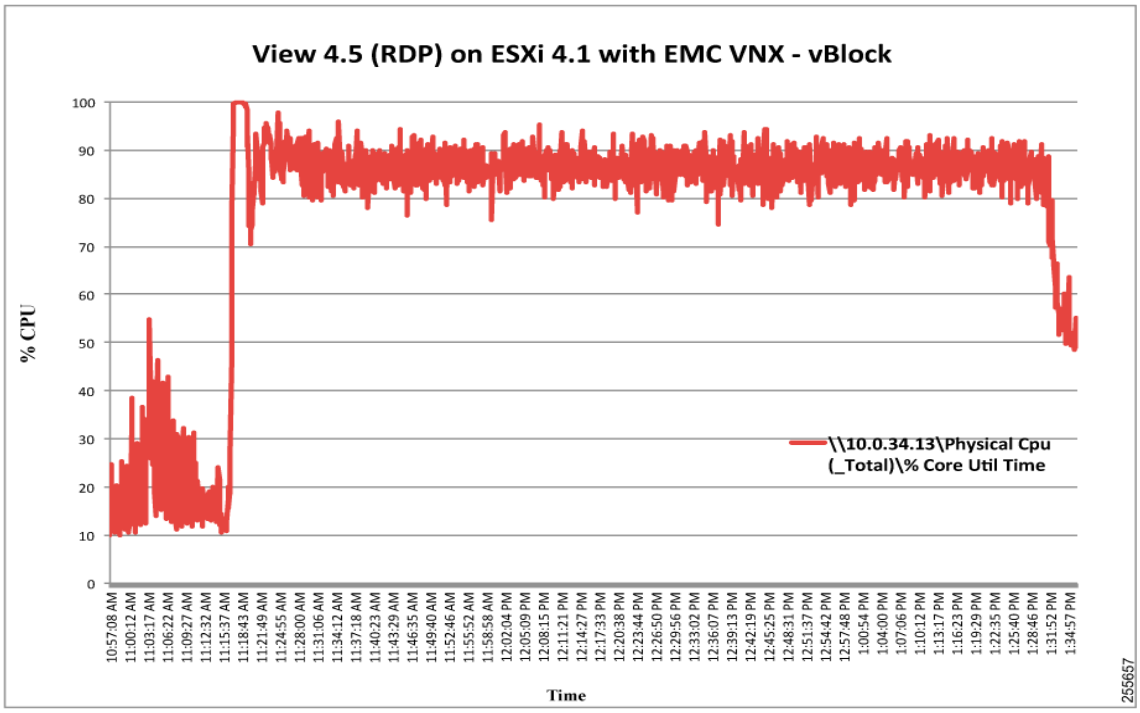


Figure 4 Memory Utilization for View 4.5 (RDP) on a UCS B250 M2 using EMC

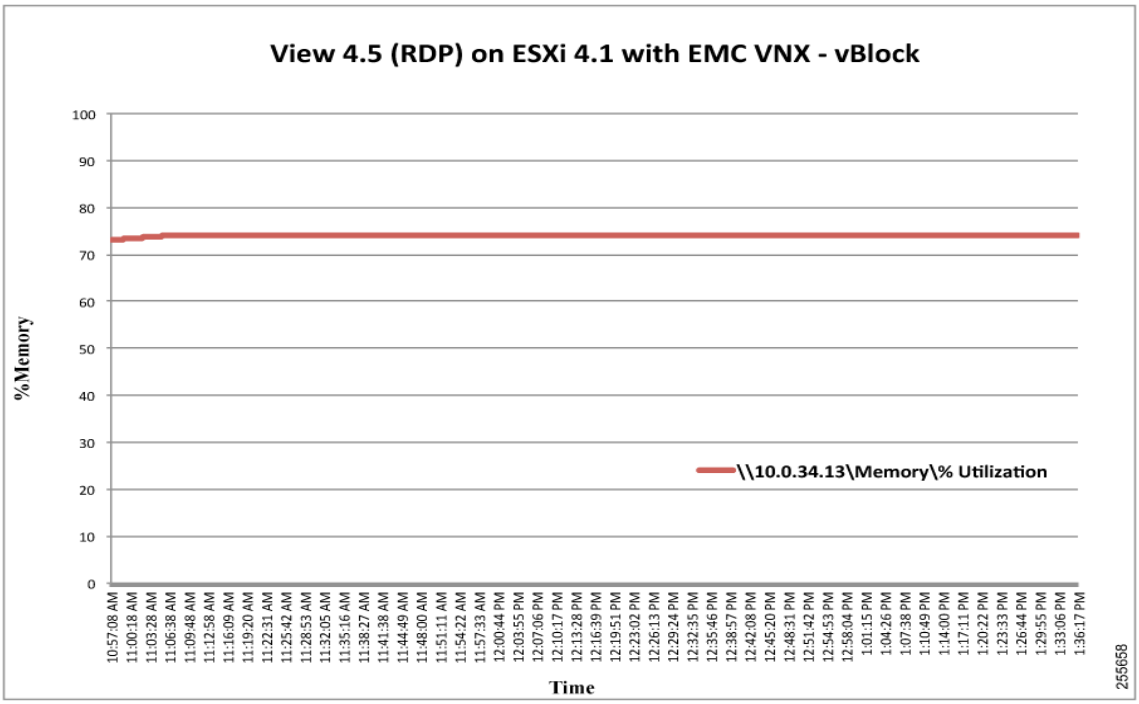


Figure 5 IOPS Measured for View 4.5 (RDP) on a UCS B250 M2 using EMC

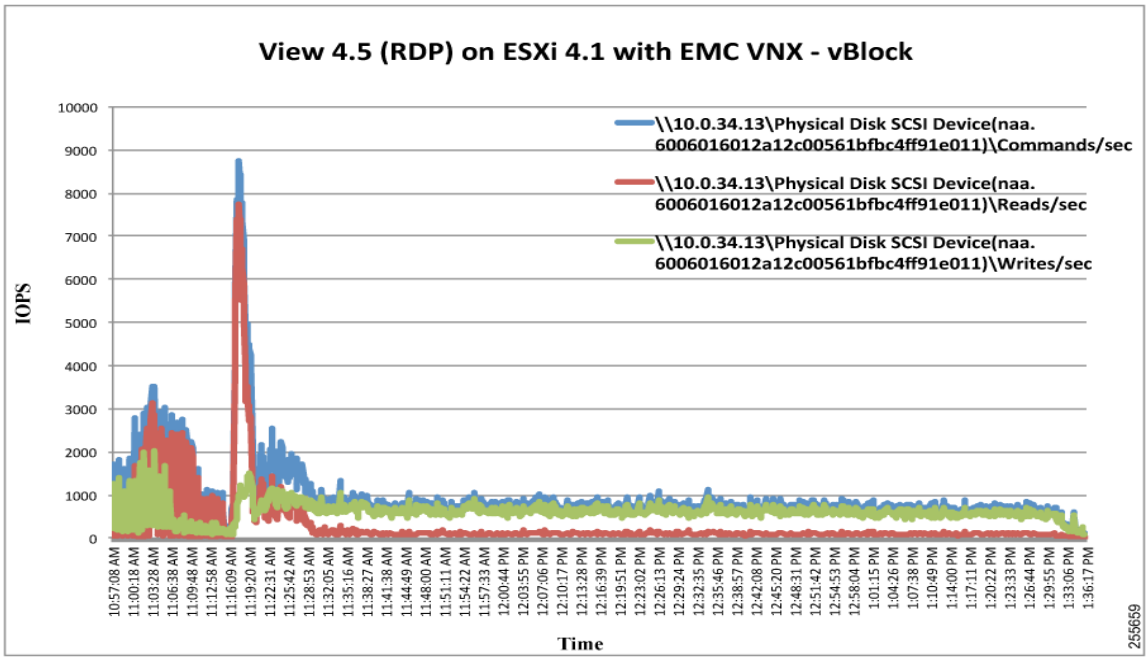


Figure 6 IO Bandwidth Measured for View 4.5 (RDP) on a UCS B250 M2 using EMC

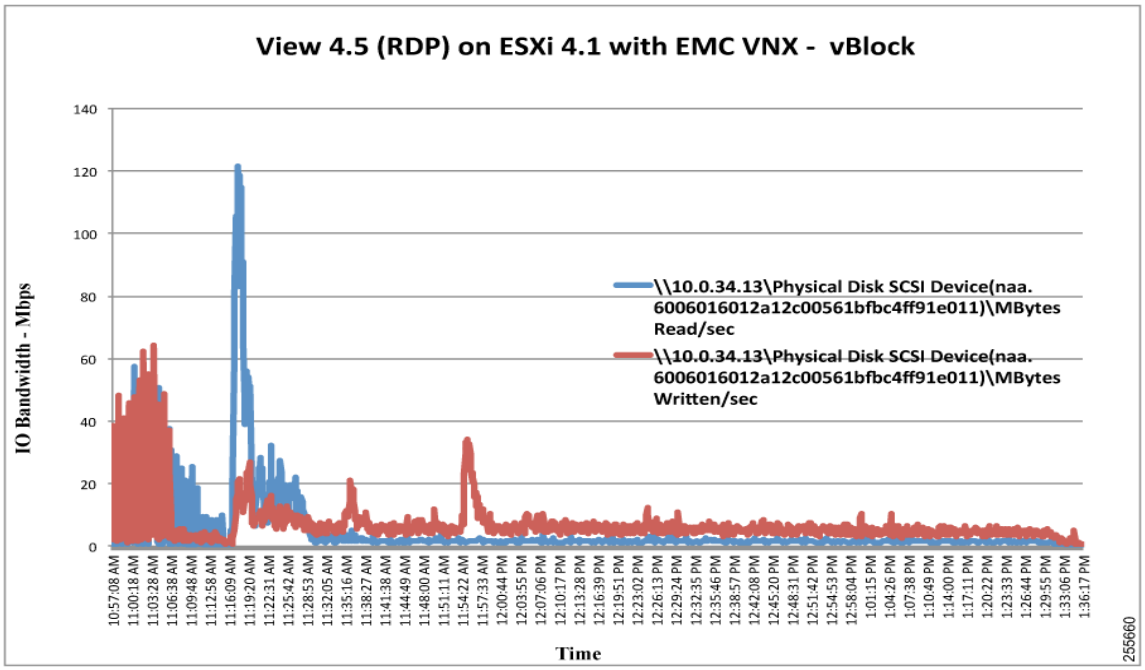


Figure 7 IO Latency Measured for View 4.5 (RDP) on a UCS B250 M2 using EMC

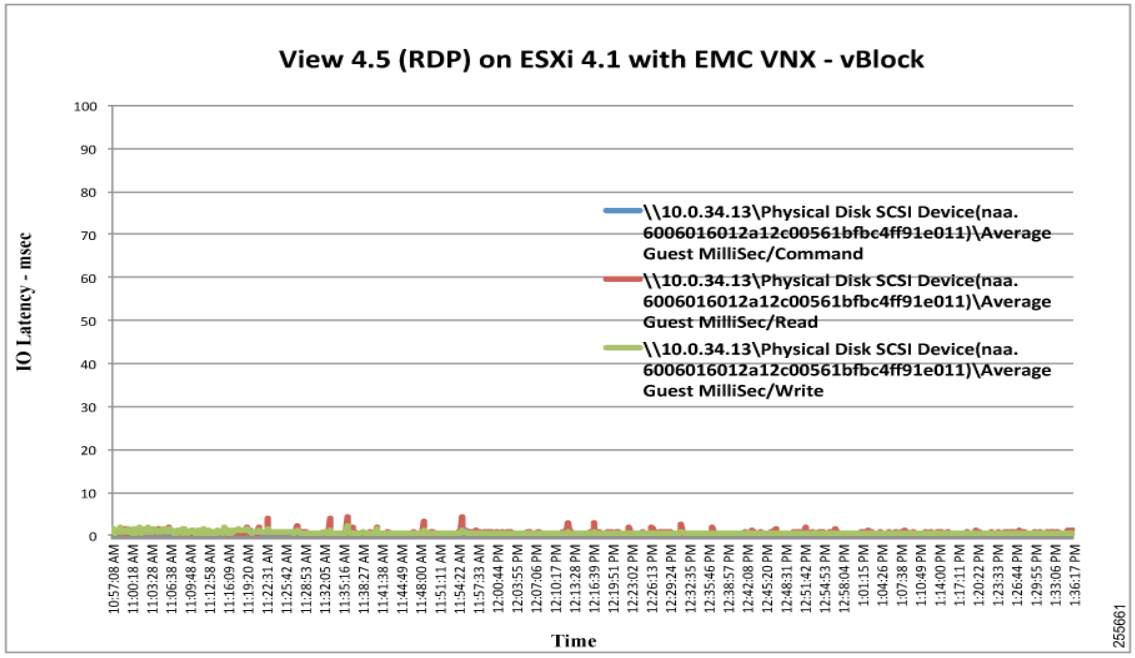


Figure 8 IO Statistics from EMC VNX 5500

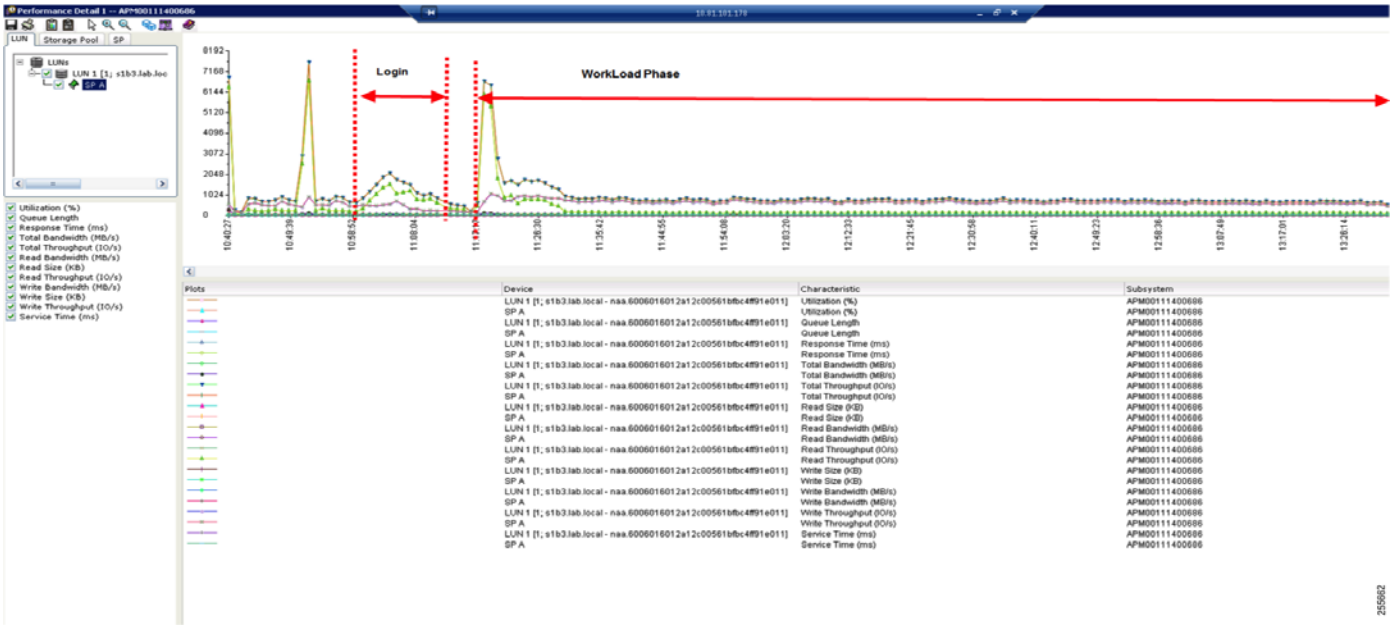
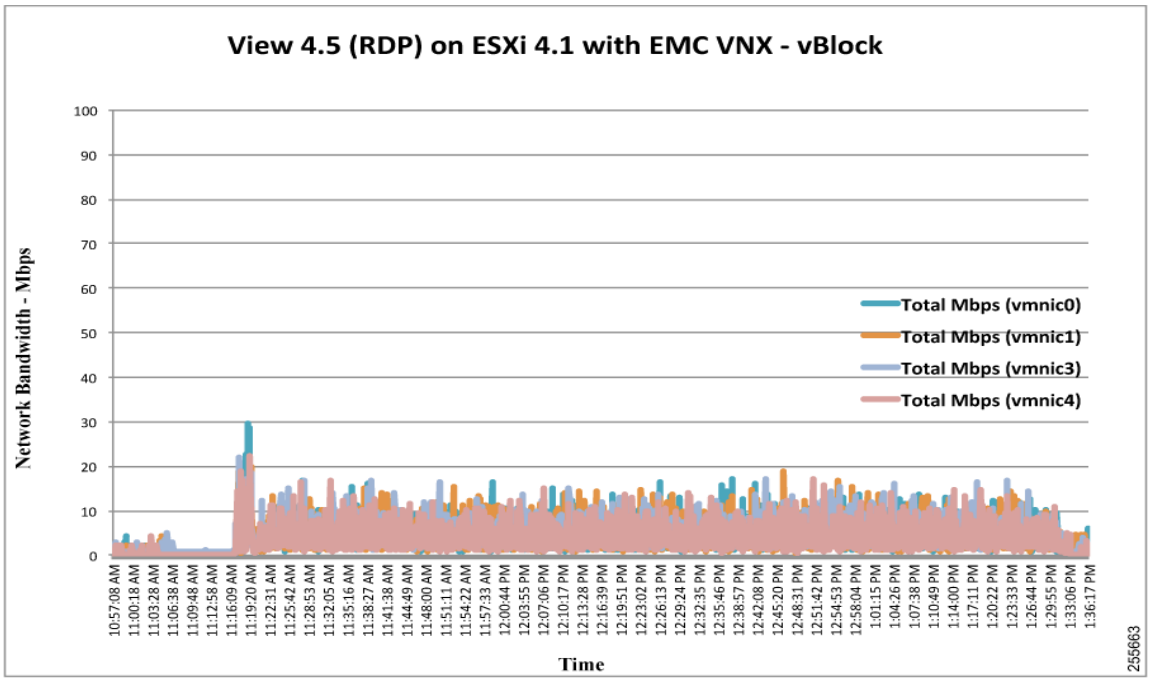


Figure 9 Network Bandwidth Measured for View 4.5 (RDP) on a UCS B250 M2 using EMC VNX 5500



FlexPod Profile

This section covers the results of the single server scalability tests done for a UCS B250 M2 using a FlexPod infrastructure environment with Windows 7 32b desktops running on View 4.5 and ESXi. Based on the testing done with Cisco KW+ workload, 80 virtual desktops can be supported on a UCS B250 M2 using this deployment profile.

Test Environment and Setup

- View 4.5 on ESXi 4.1; PCoIP
- HVD Profile:
  - Windows 7 32b with 1.5G of memory and 20G of disk space
  - 1 vCPU, Persistent desktop
- Workload Profile: Cisco KW+ (Cisco Unified Personal Communicator 8.5 in deskphone mode, IE, Microsoft Office 2007 Apps, Acrobat) with optimized antivirus solution from a leading vendor
- UCS Server: B250 M2 with 192 G of memory - Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz and 1GE uplinks
- Storage: NFS on NetApp FAS 3170 with PAM2 (512G of cache)
- All of the data shown in the graph below is collected using resxtop with a polling interval of 5s except for
- User Experience is measured subjectively by spot-checking workloads running on the sessions
- For this profile, data is captured and graphed only for the workload phase

Summary of Test Results

For the deployment profile detailed in the Test Environment and Setup section above, 80 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = 70%
- Storage
  - IOPS = Write IOPS stay at an average of ~800 IOPS with reads being significantly less
  - IO Bandwidth = Avg. Write BW of ~50MBps during steady state
  - IO Latency = Peak Read & Write Latency of ~40ms seen during workload start
- Network Bandwidth Utilization = < 20Mbps during steady state

Figure 10 CPU Utilization for View 4.5 (PCoIP) on a UCS B250 M2 using NetApp

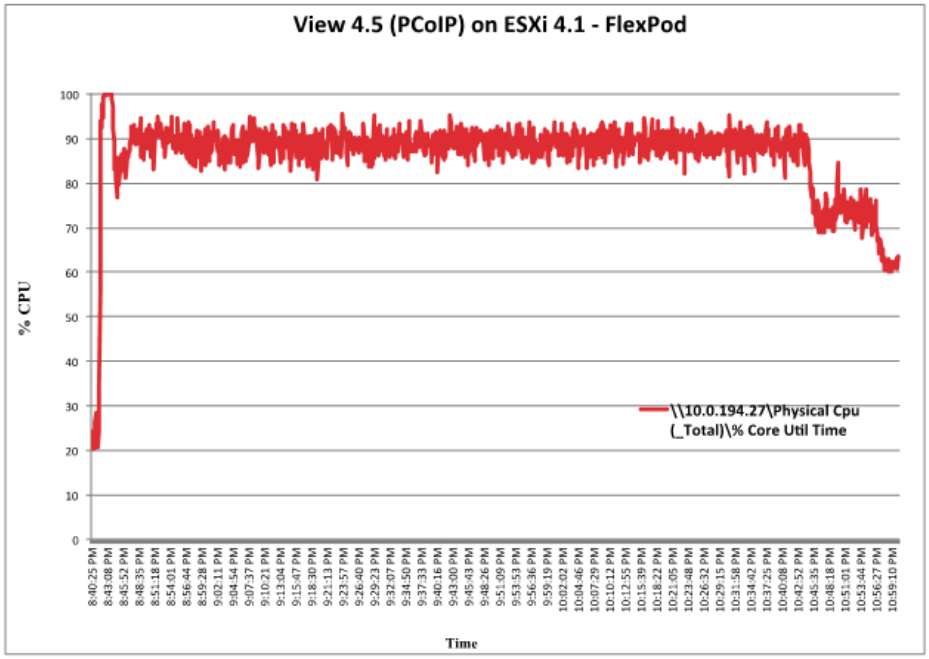


Figure 11 Memory utilization for View 4.5 (PCoIP) on a UCS B250 M2 using NetApp

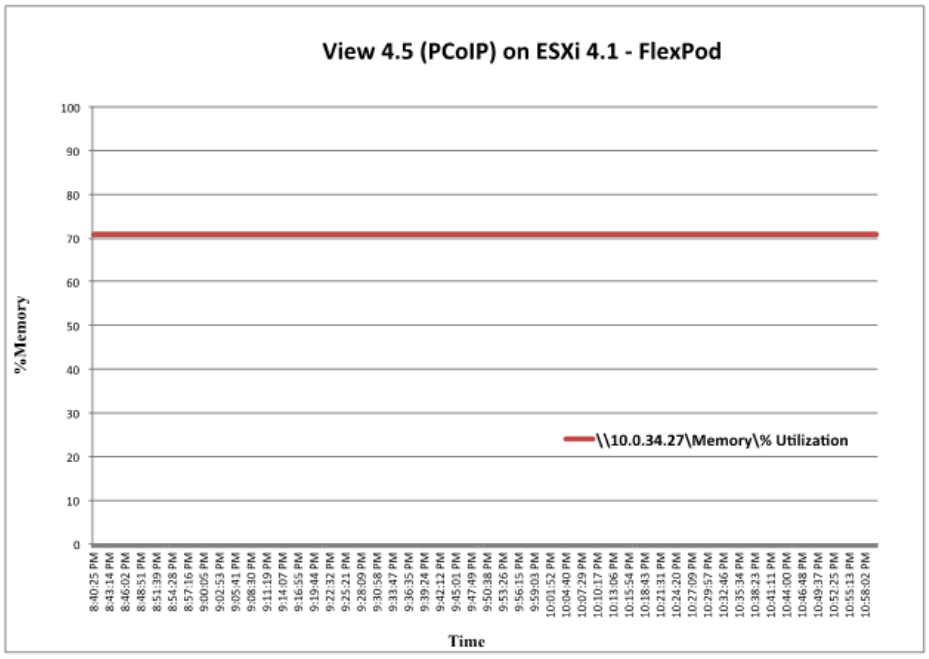


Figure 12 IOPS Measured for View 4.5 (PCoIP) on a UCS B250 M2 using NetApp

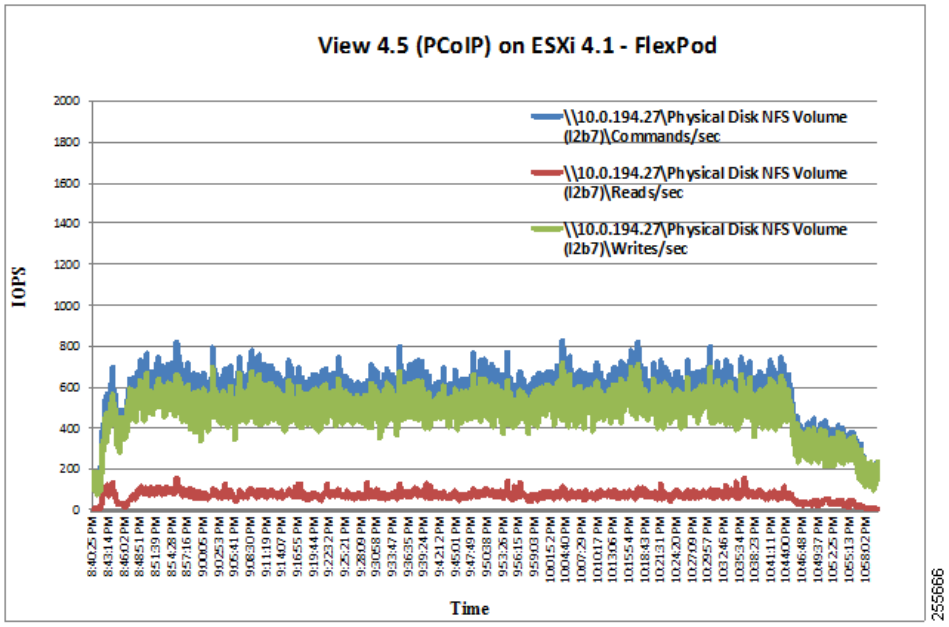


Figure 13 IO Bandwidth Measured for View 4.5 (PCoIP) on a UCS B250 M2 using NetApp

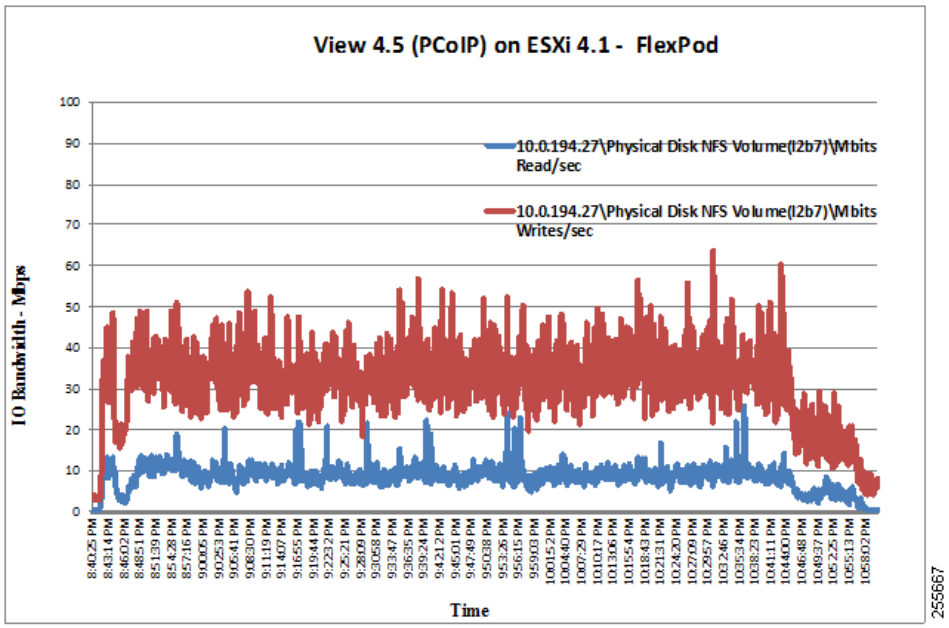


Figure 14 IO Latency Measured for View 4.5 (PCoIP) on a UCS B250 M2 using NetApp

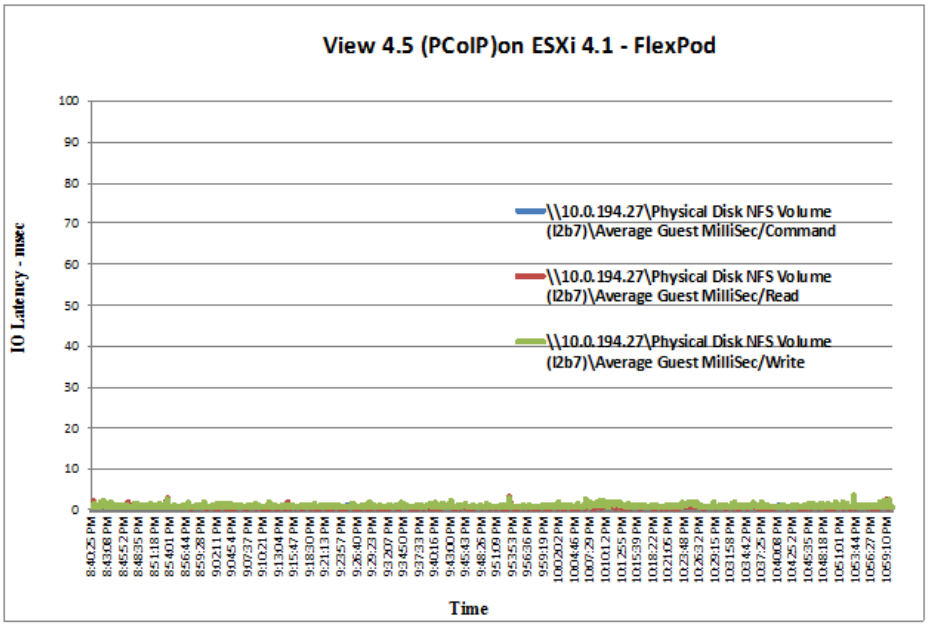
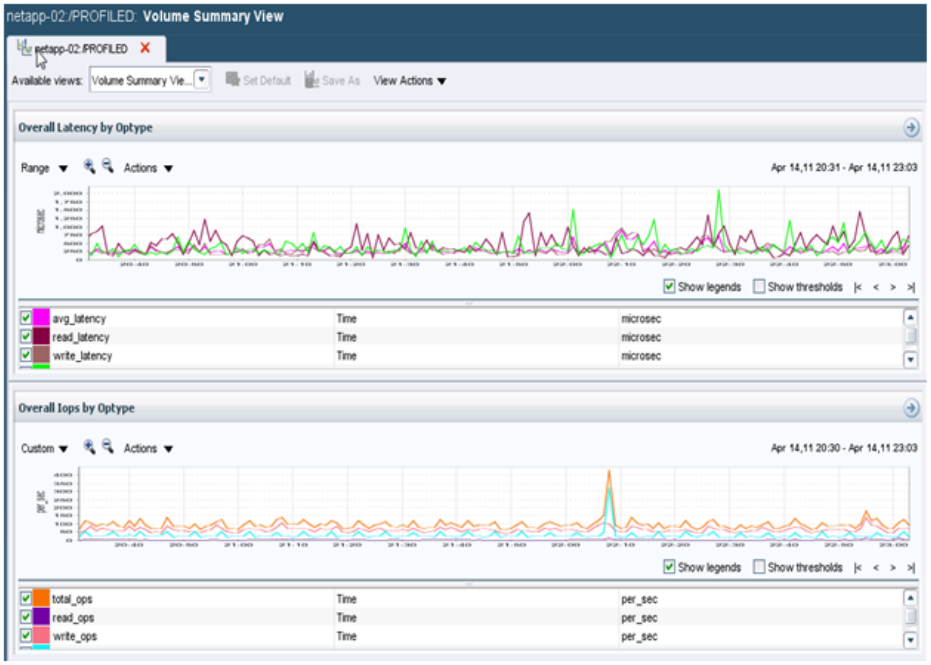


Figure 15 IO Statistics from NetApp



View4.6/ESXi4.1/RDP/B250M2 Profile –New CPU Counter

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 across a FlexPod infrastructure with Windows 7 32b desktops running on View 4.6 and ESXi 4.1 U1. Results indicate that ~115 virtual desktops can be supported on a UCS B250 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are CPU bound for this profile.

Test Profile

Desktop Virtualization

- VMware View 4.6
- Connection Protocol – RDP
- Linked Clones

Hypervisor

VMware ESXi 4.1 U1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktop

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- 192G RAM (16 x 8GB DIMMS)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (ver1.6)

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

- NAS - NFS
- NetApp FAS 3170 with PAM2 (512G of cache)

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s

- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login and Workload phases

Summary of Test Results

For the deployment profile detailed above, 115 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = ~95%
- Application Response times – Success Criteria met

Performance Charts

Figure 16 CPU Utilization Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

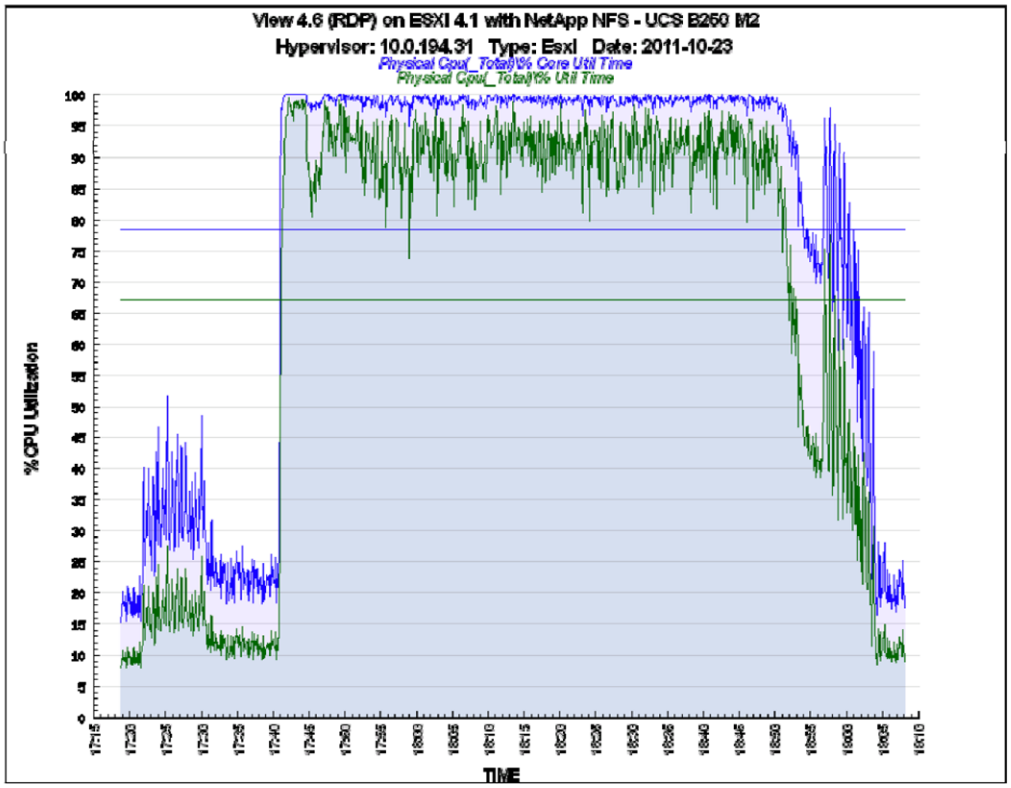


Figure 17 Memory Utilization Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

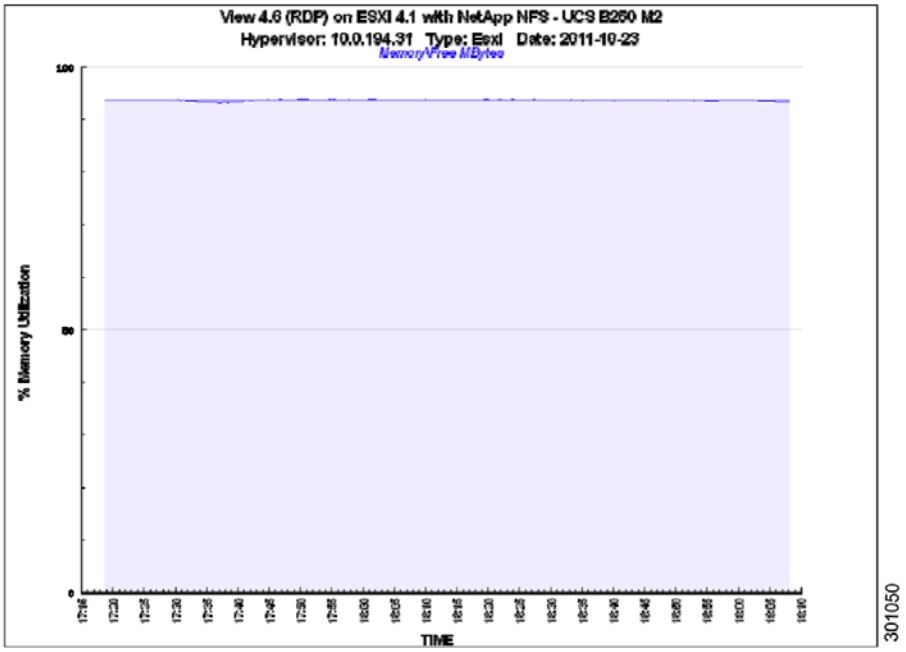


Figure 18 IO Statistics Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

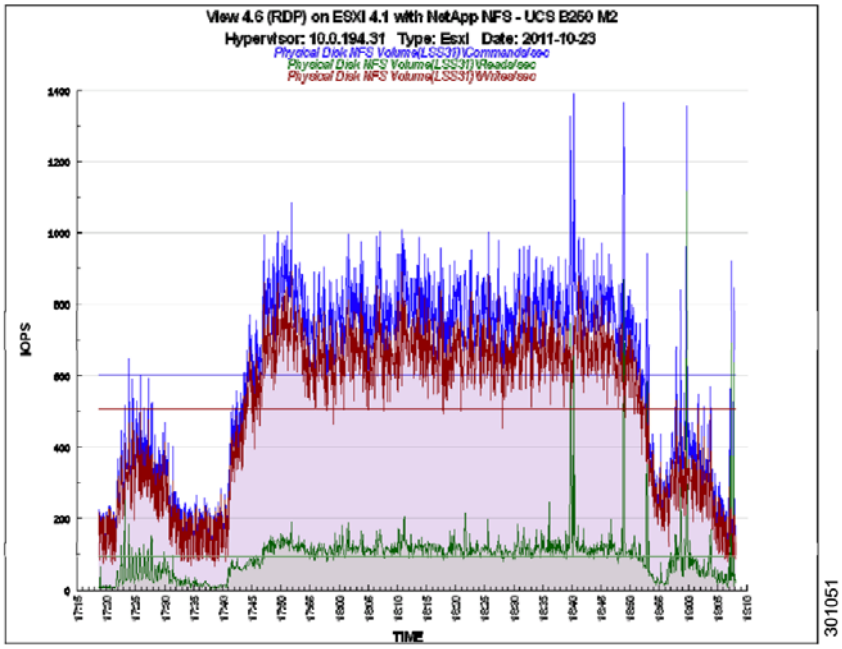


Figure 19 IO Latency Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

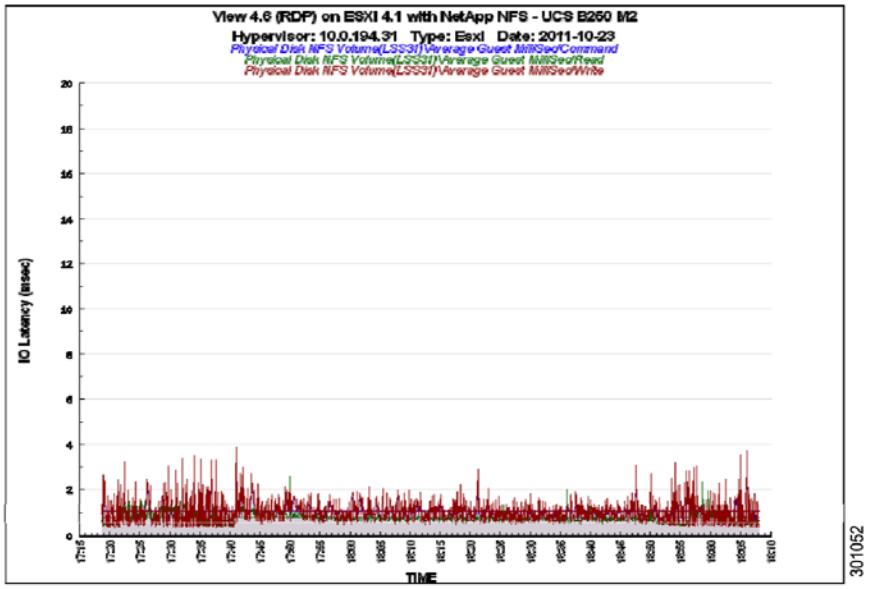


Figure 20 IO BW Utilization Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

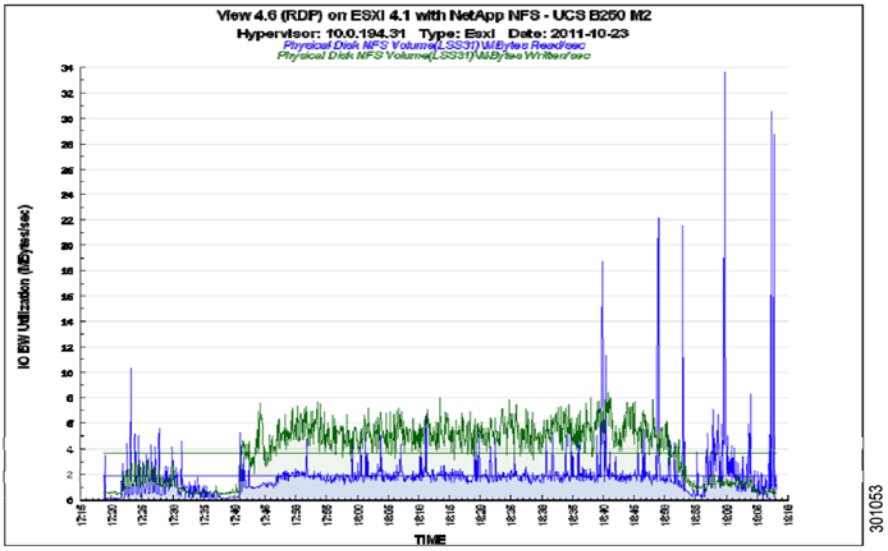
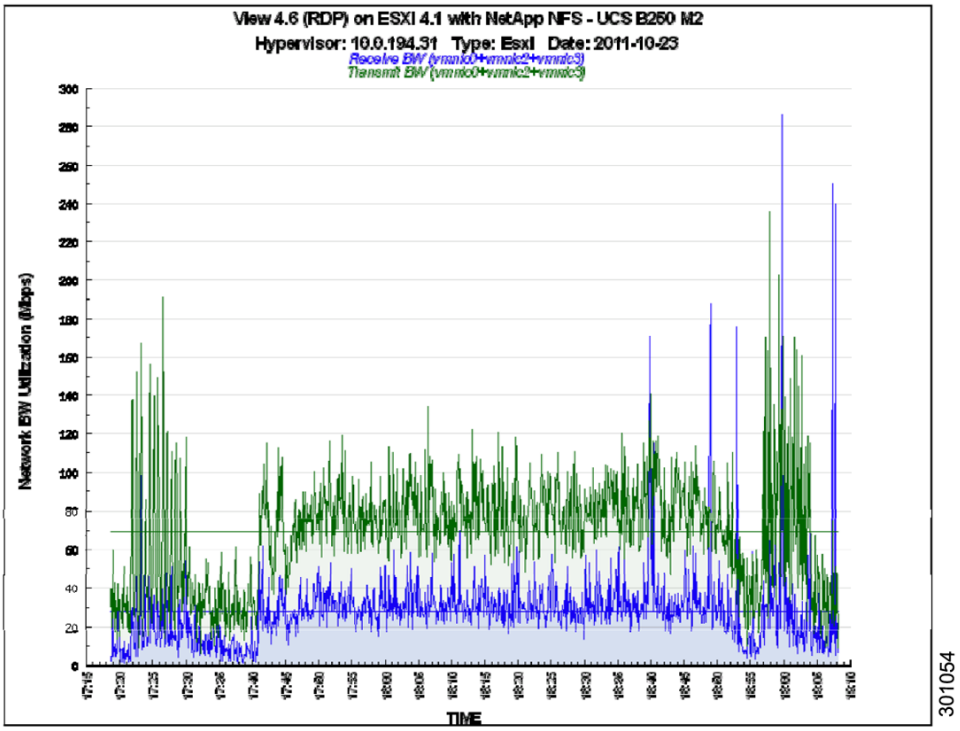


Figure 21 Network BW Utilization Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 7 Response Times for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test(sec)
Cisco Unified Personal Communicator 8.5 in deskphone control mode	5s	1.3s
Outlook	5s	3.3s
Word	10s	7.5s
Excel	5s	0.95s
Powerpoint	5s	0.7s
Internet Explorer	5s	3.7s
Acrobat	5s	0.6s

View4.6/ESXi4.1/PCoIP/B250M2 Profile –New CPU Counter

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 across a FlexPOD infrastructure with Windows 7 32b desktops running on View 4.6 and ESXi 4.1 U1. Results indicate that ~95 virtual desktops can be supported on a UCS B250 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are CPU bound for this profile.

Test Profile

Desktop Virtualization

- VMware View 4.6
- Connection Protocol – PCoIP
- Linked Clones

Hypervisor

VMware ESXi 4.1 U1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktop

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- 192G RAM (16 x 8GB DIMMS)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (ver1.6)

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

- NAS - NFS
- NetApp FAS 3170 with PAM2 (512G of cache)

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s

- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login and Workload phases

Summary of Test Results

For the deployment profile detailed above, 95 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = ~80%
- Application Response times – Success Criteria met

Performance Charts

Figure 22 CPU Utilization Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

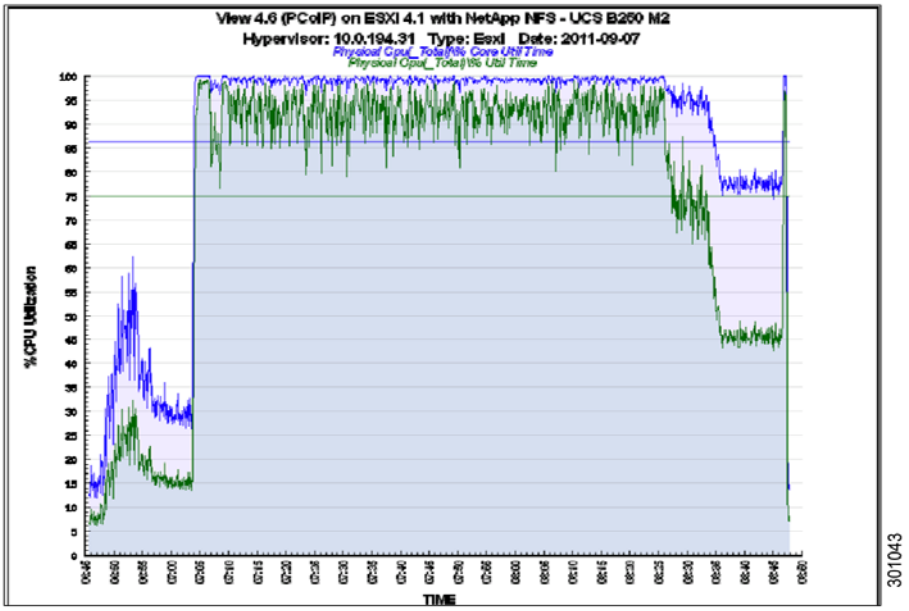


Figure 23 CPU Utilization Chart for View4.6/ESXi4.1/PCoIP/B250M2/NetApp Profile

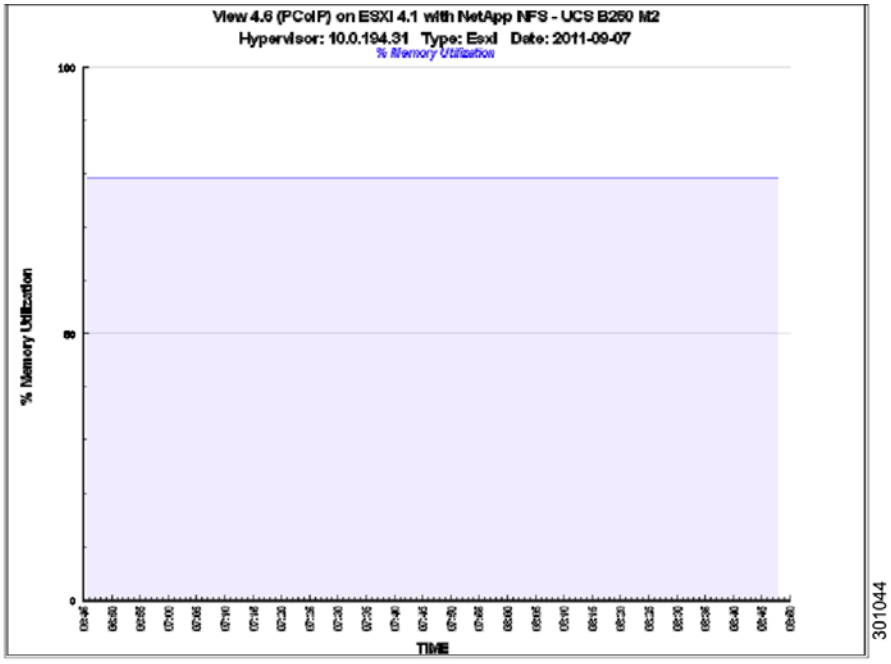


Figure 24 IO Statistics Chart for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

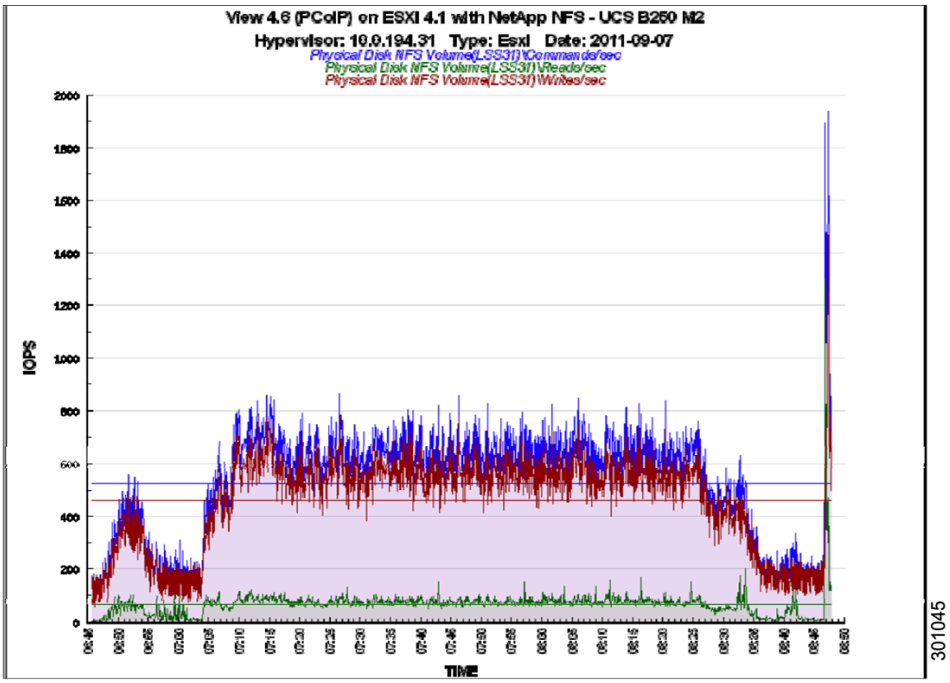


Figure 25 Latency Chart for View4.6/ESXi4.1/PCoIP/B250M2/NetApp Profile

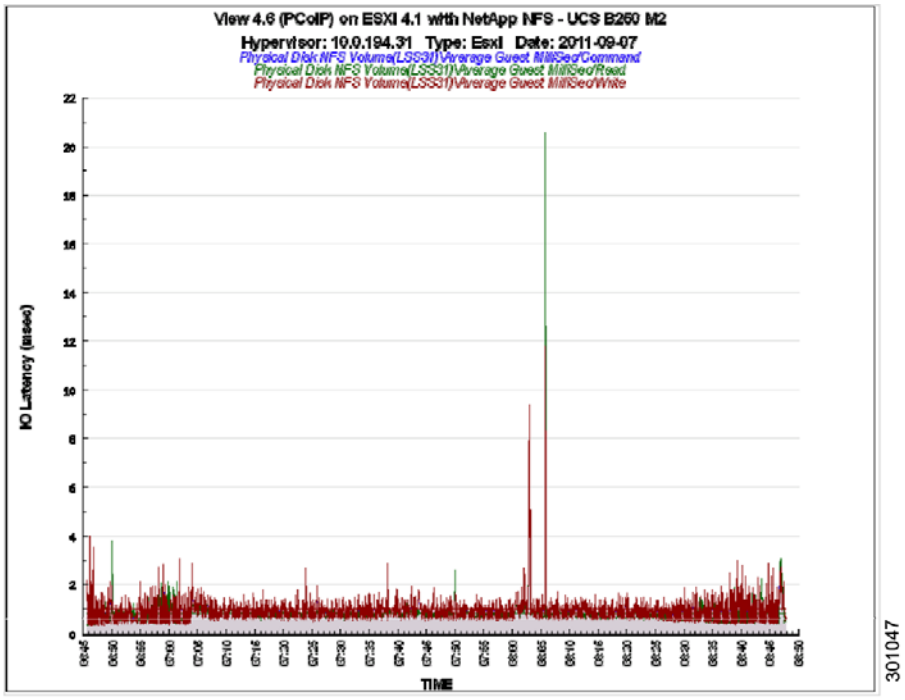


Figure 26 IO BW Utilization Chart for View4.6/ESXi4.1/PCoIP/B250M2/NetApp Profile

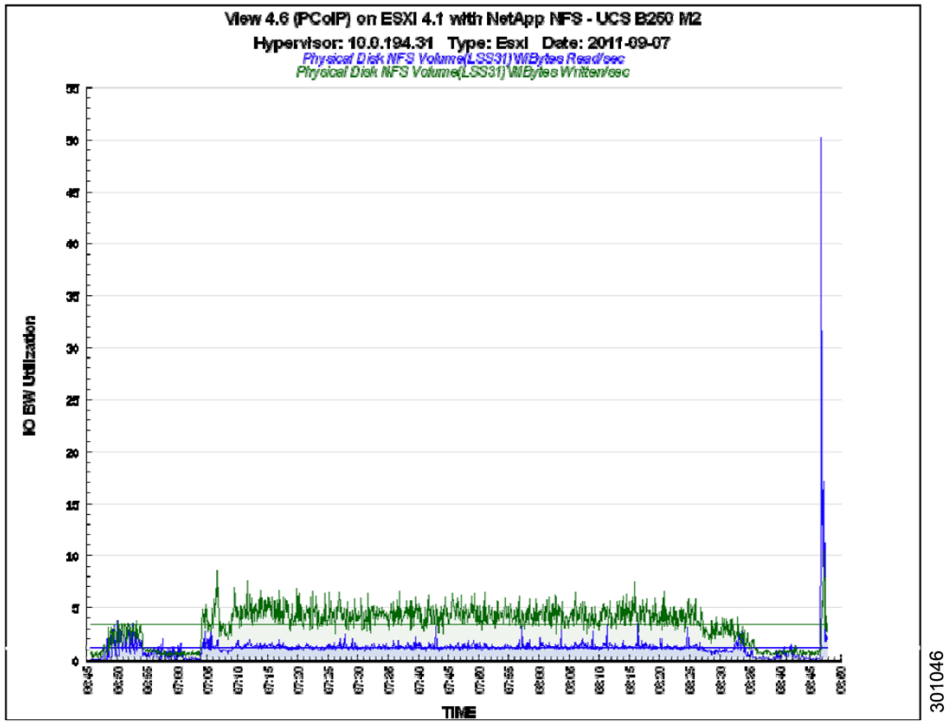
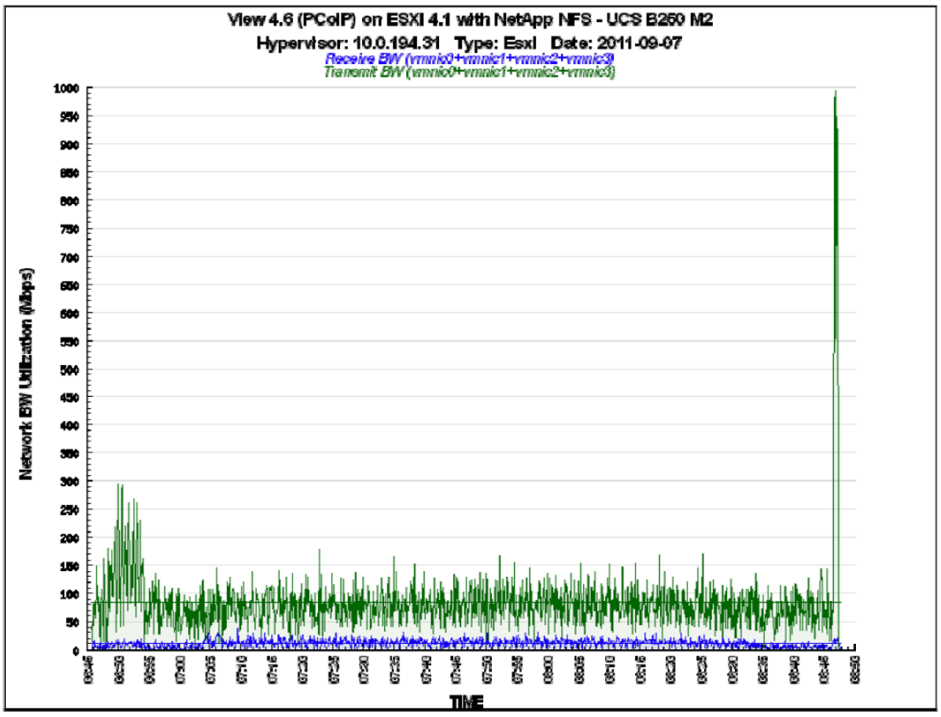


Figure 27 Network BW Utilization Chart for View4.6/ESXi4.1/PCoIP/B250M2/NetApp Profile



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 8 Response Times for View4.6/ESXi4.1/RDP/B250M2/NetApp Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test(sec)
Cisco Unified Personal Communicator 8.5 in deskphone control mode	5s	1.4s
Outlook	5s	3.1s
Word	5s	7.2s
Excel	5s	0.95s
Powerpoint	5s	0.73s
Internet Explorer	5s	3.5s
Acrobat	5s	0.58s

View5/ESXi5/RDP/B250M2 Profile - Vblock

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 across a Vblock infrastructure with Windows 7 32b desktops running on View 5 and ESXi 5.0. Results indicate that ~125 virtual desktops can be supported on a UCS B250 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are CPU bound for this profile.

Test Profile

Desktop Virtualization

- VMware View 5.0
- Connection Protocol – RDP
- Linked Clones

Hypervisor

VMware ESXi 5.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktop

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- 192G RAM (16 x 8GB DIMMS)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (ver2.5)

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

- Fibre Channel attached SAN
- VSPEX (EMC VNX 5500)

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s

- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login and Workload phases

Summary of Test Results

For the deployment profile detailed above, 125 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = ~60%
- Application Response times – Success Criteria met

Performance Charts

Figure 28 CPU Utilization Chart for View5/ESXi5/RDP/B250M2/EMC Profile

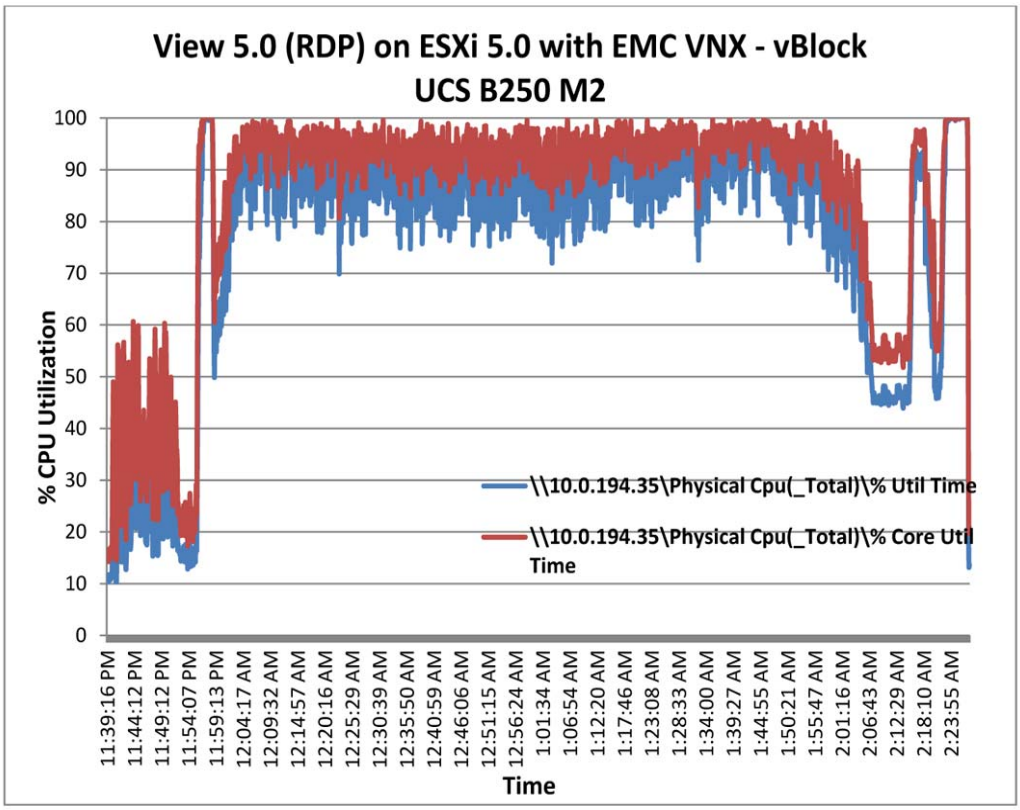


Figure 29 Memory Utilization Chart for View5/ESXi5/RDP/B250M2/EMC Profile

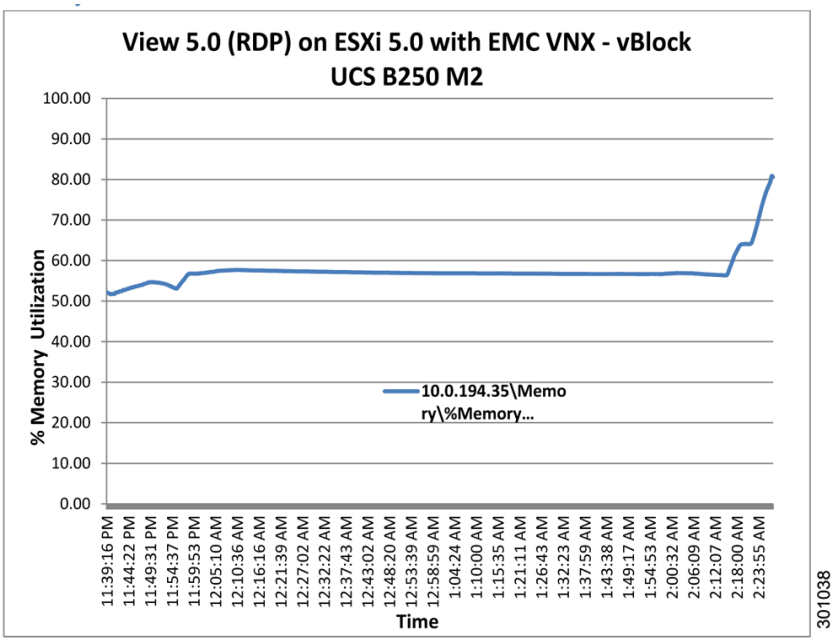


Figure 30 IOPS Chart for View5/ESXi5/RDP/B250M2/EMC Profile

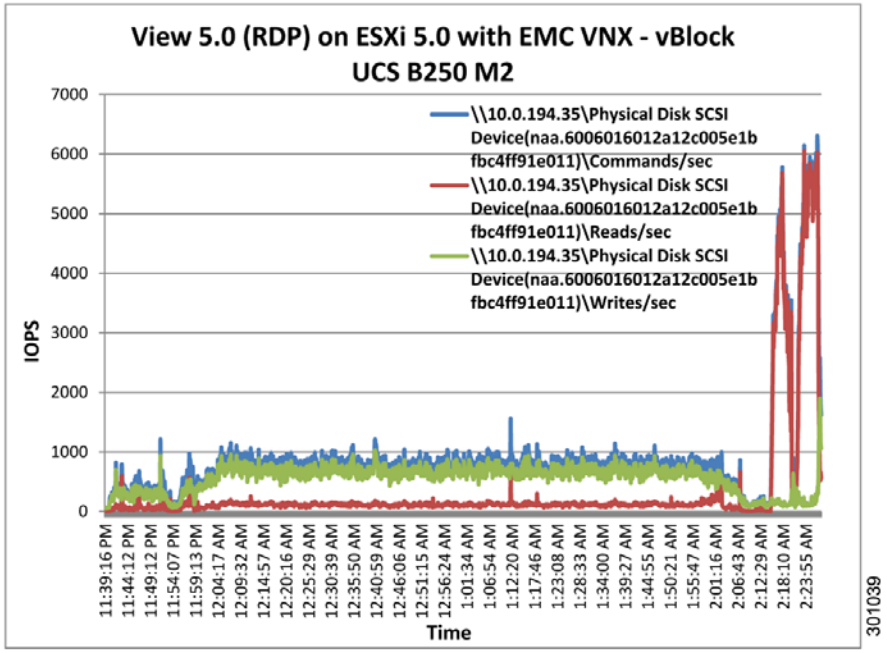


Figure 31 IO Latency Chart for View5/ESXi5/RDP/B250M2/EMC Profile

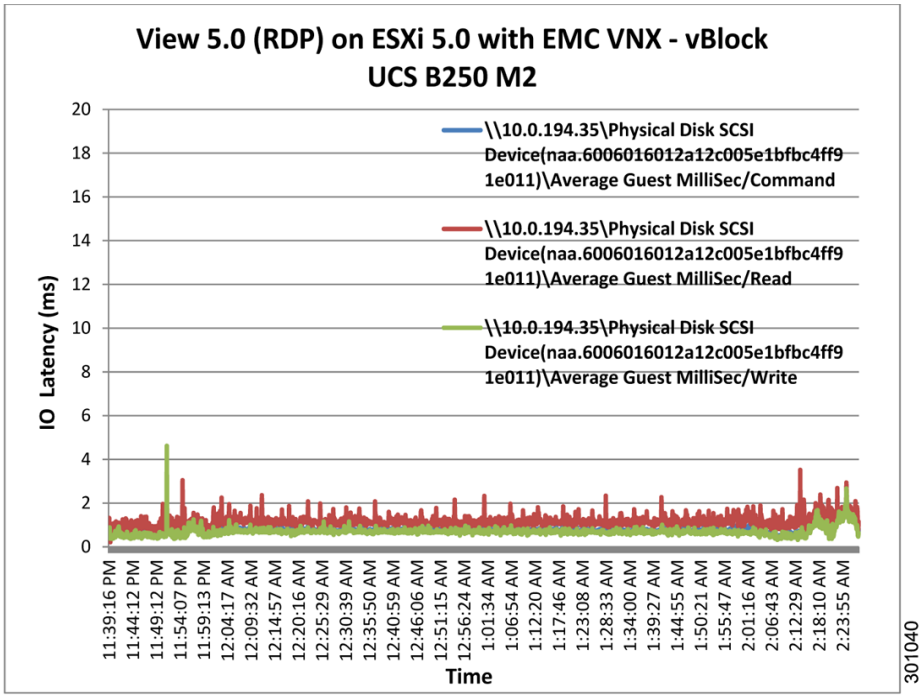


Figure 32 IO BW Utilization Chart for View5/ESXi5/RDP/B250M2/EMC Profile

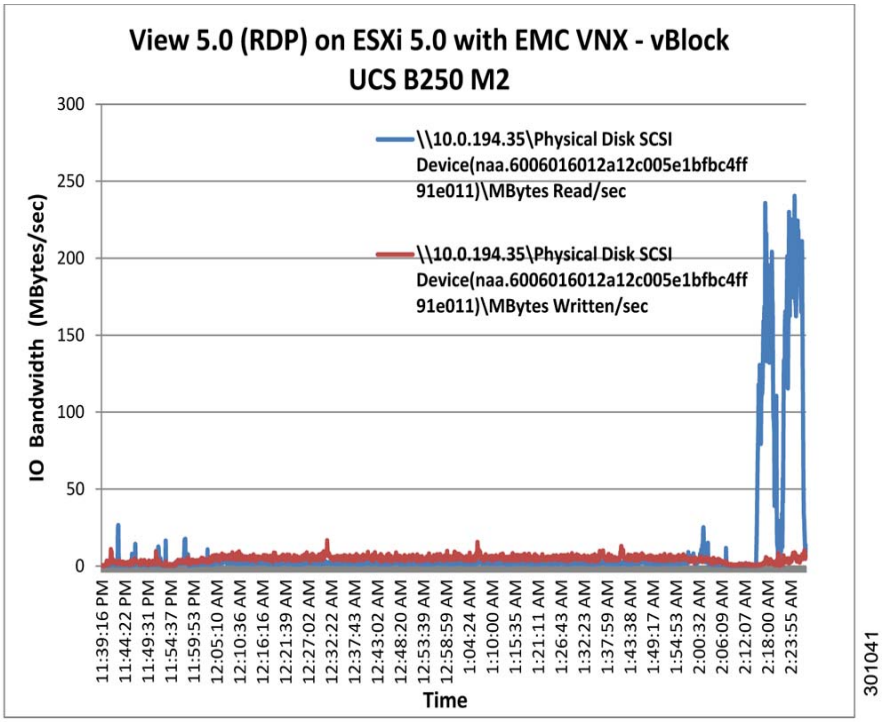
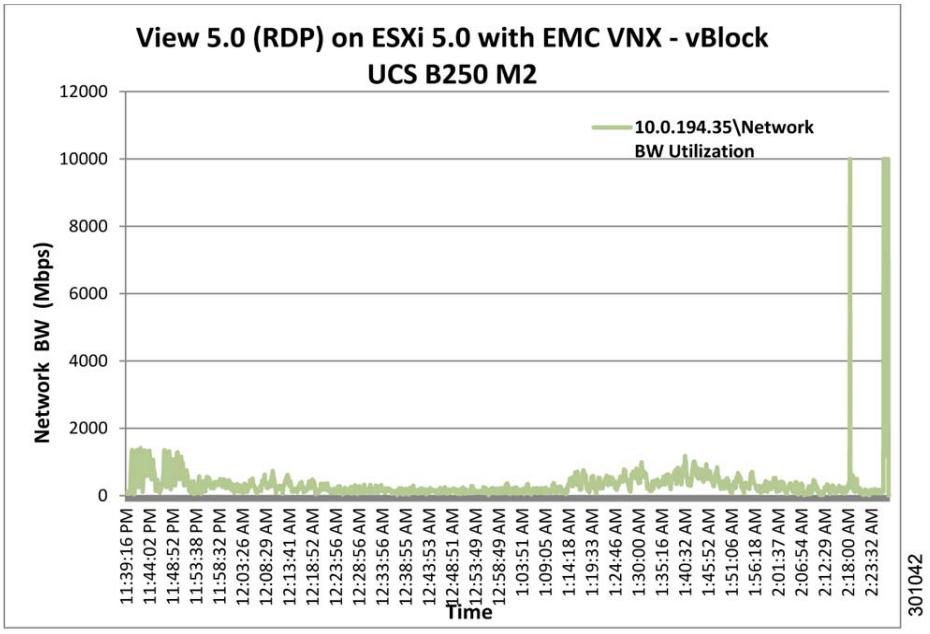


Figure 33 Network BW Utilization Chart for View5/ESXi5/RDP/B250M2/EMC Profile



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 9 Response Times for View5.0/ESXi5.0/RDP/B250M2/EMC Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test(sec)
Cisco Unified Personal Communicator 8.5 in deskphone control mode	5s	1.6s
Outlook	5s	2.2s
Word	5s	0.7s
Excel	5s	0.8s
Powerpoint	5s	0.6s
Internet Explorer	5s	0.8s
Acrobat	5s	0.5s

View5/ESXi5/PCoIP/B250M2 Profile - Vblock

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 across a Vblock infrastructure with Windows 7 32b desktops running on View 5 and ESXi 5.0. Results indicate that ~115 virtual desktops can be supported on a UCS B250 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are CPU bound for this profile.

Test Profile

Desktop Virtualization

- VMware View 5.0
- Connection Protocol – PCoIP
- Linked Clones

Hypervisor

VMware ESXi 5.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktop

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- 192G RAM (16 x 8GB DIMMS)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (ver2.5)

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

- Fibre Channel attached SAN
- VSPEX (EMC VNX 5500)

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login and Workload phases

Summary of Test Results

For the deployment profile detailed above, 115 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = ~60%
- Application Response times – Success Criteria met

Performance Charts

Figure 34 CPU Utilization Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile

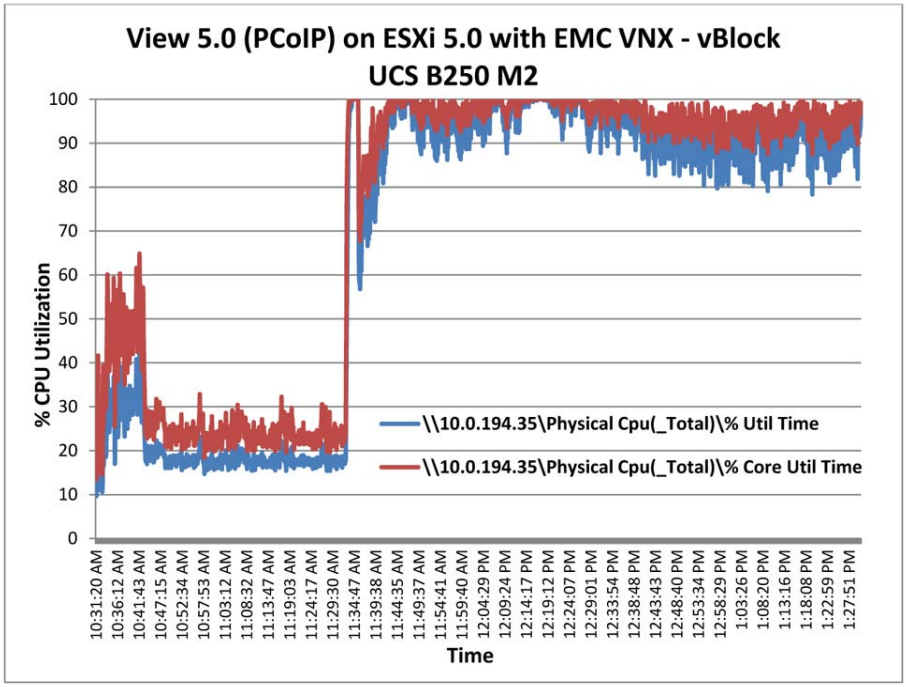


Figure 35 Memory Utilization Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile

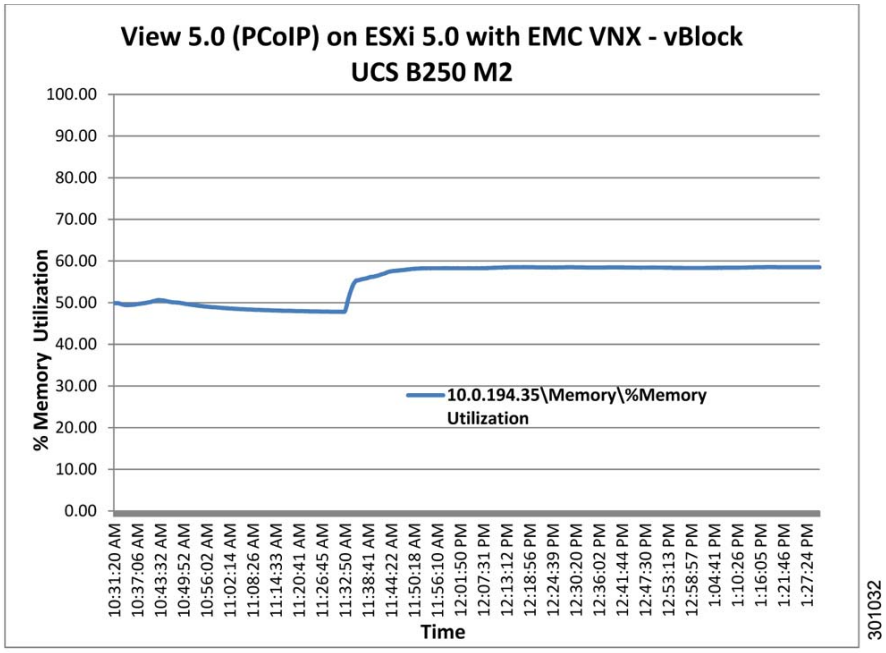


Figure 36 IOPS Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile

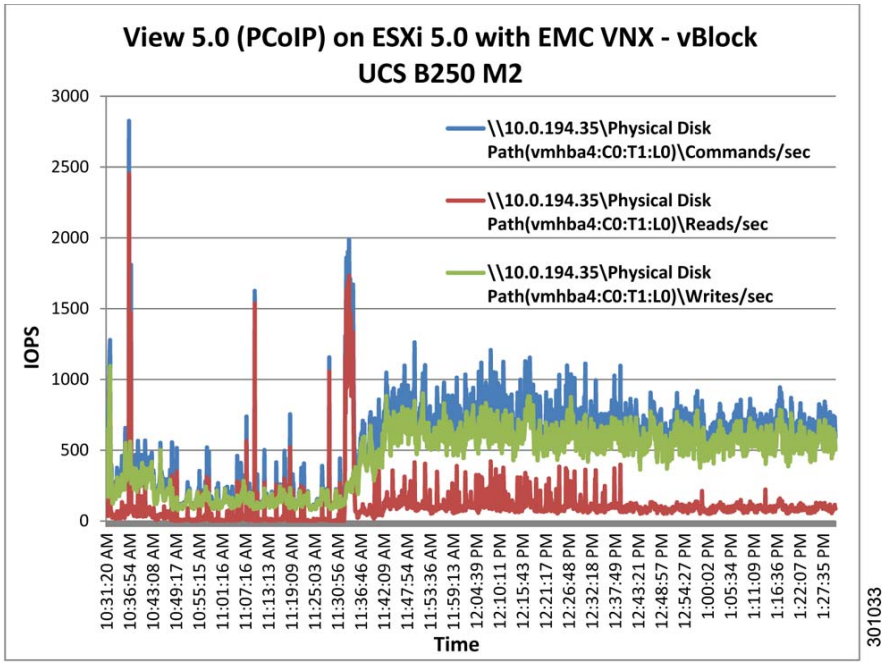


Figure 37 IO Latency Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile

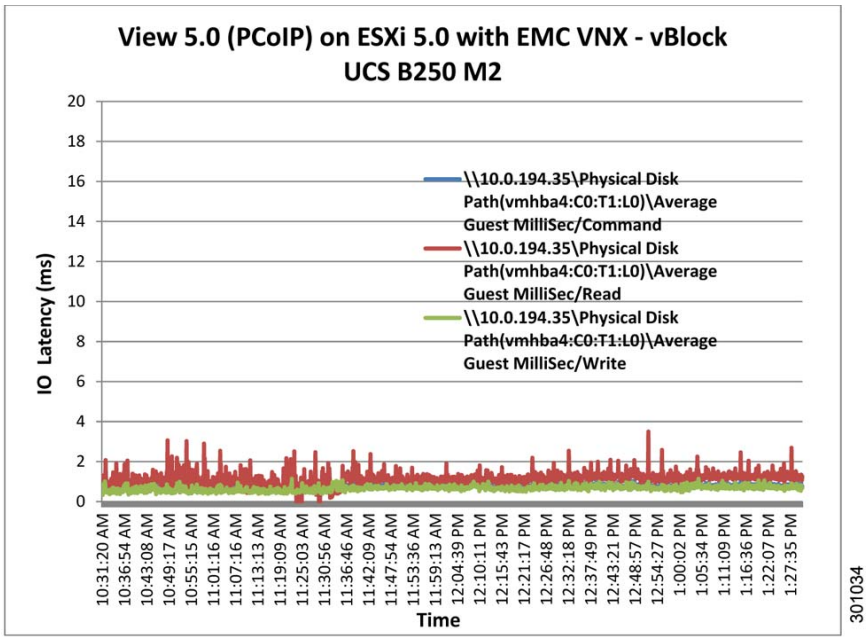
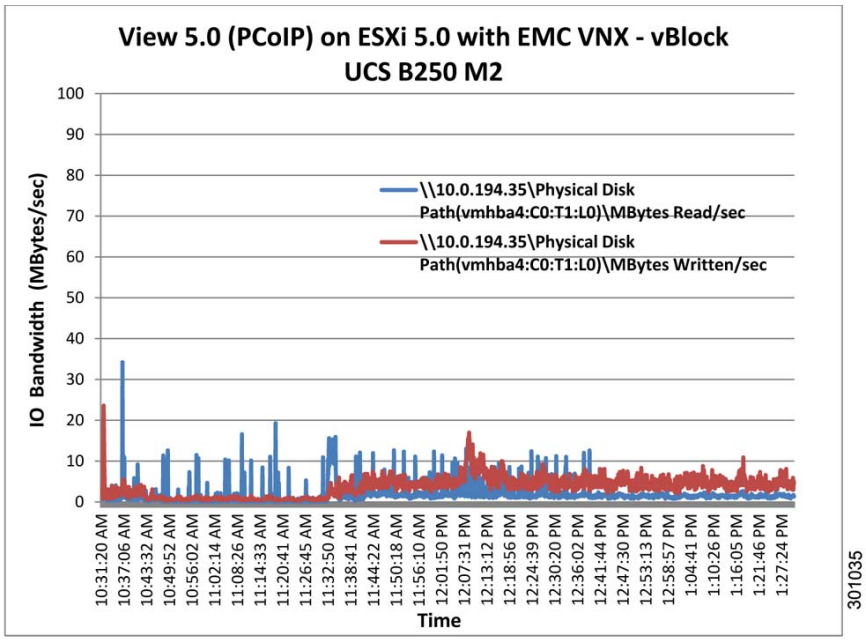
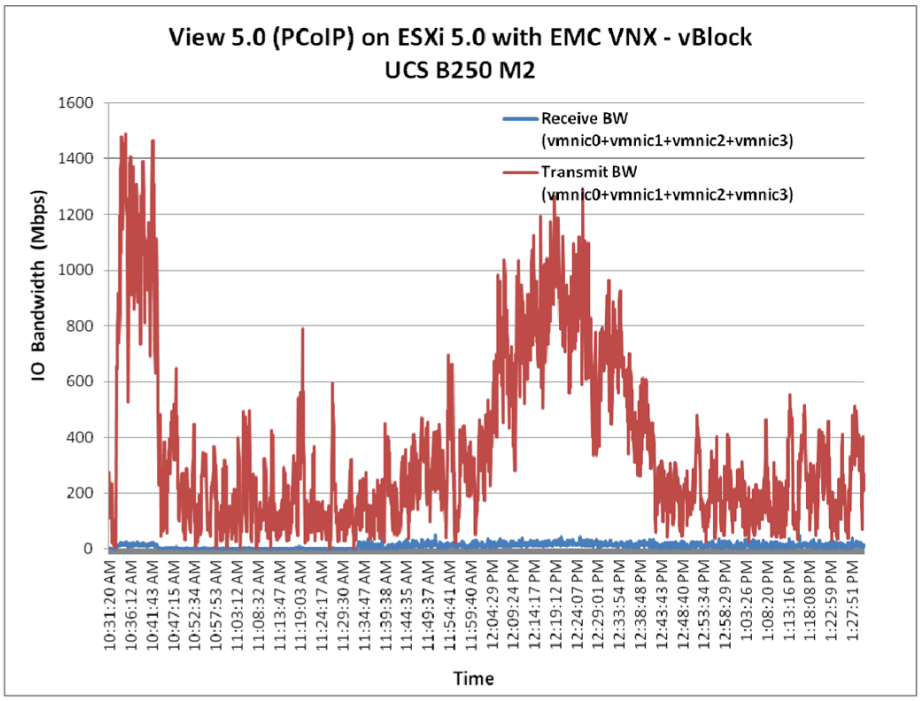


Figure 38 IO BW Utilization Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile



Network Bandwidth Utilization

Figure 39 Network BW Utilization Chart for View5/ESXi5/PCoIP/B250M2/EMC Profile



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 10 Response Times for View5.0/ESXi5.0/RDP/B250M2/EMC Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test(sec)
Cisco Unified Personal Communicator 8.5 in deskphone control mode	5s	2.4s
Outlook	5s	4.0s
Word	5s	2.4s
Excel	5s	2.6s
Powerpoint	5s	1.6s
Internet Explorer	5s	2.5s
Acrobat	5s	1.5s

HVD Scalability for View5/ESXi5.0 Profile on UCS B230 M2

This section provides the detailed results of the single server scale and performance tests done for a UCS B230 M2 across a FlexPod infrastructure with Windows 7 32b desktops running on View 5 and ESXi 5.0. Results indicate that ~160 virtual desktops can be supported on a UCS B230 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are memory bound for this profile.

Detailed Performance Results

This section provides a detailed overview of the results based on the testing done in the end-to-end Cisco Virtual Workspace system using a Cisco KW+ workload.

Summary of Test Results

For the deployment profile detailed above, 160 VMs can be supported on a Cisco UCS B230 M2 with the following performance metrics.

Average CPU Utilization = ~90% (Steady state)

Average Memory Utilization = ~95% with negligible transparent page sharing (<=1%)

Application Response times – Success Criteria met

Average IO Latency <20ms (Actual = <10ms)

Test Profile

Desktop Virtualization

- VMware View 5
- Connection Protocol – PCoIP
- Linked Clones

Hypervisor

VMware ESXi 5.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktops

Server Specifications

- Cisco UCS B230 M2
- Two Ten Core Intel Xeon E7-2870 processors @ 2.40 GHz
- 256G RAM (32 x 8GB DIMMS @ 1066 MHz)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (ver2.5)

- Microsoft Office 2007 Applications

- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode
- Optimized antivirus solution from a leading vendor

Storage

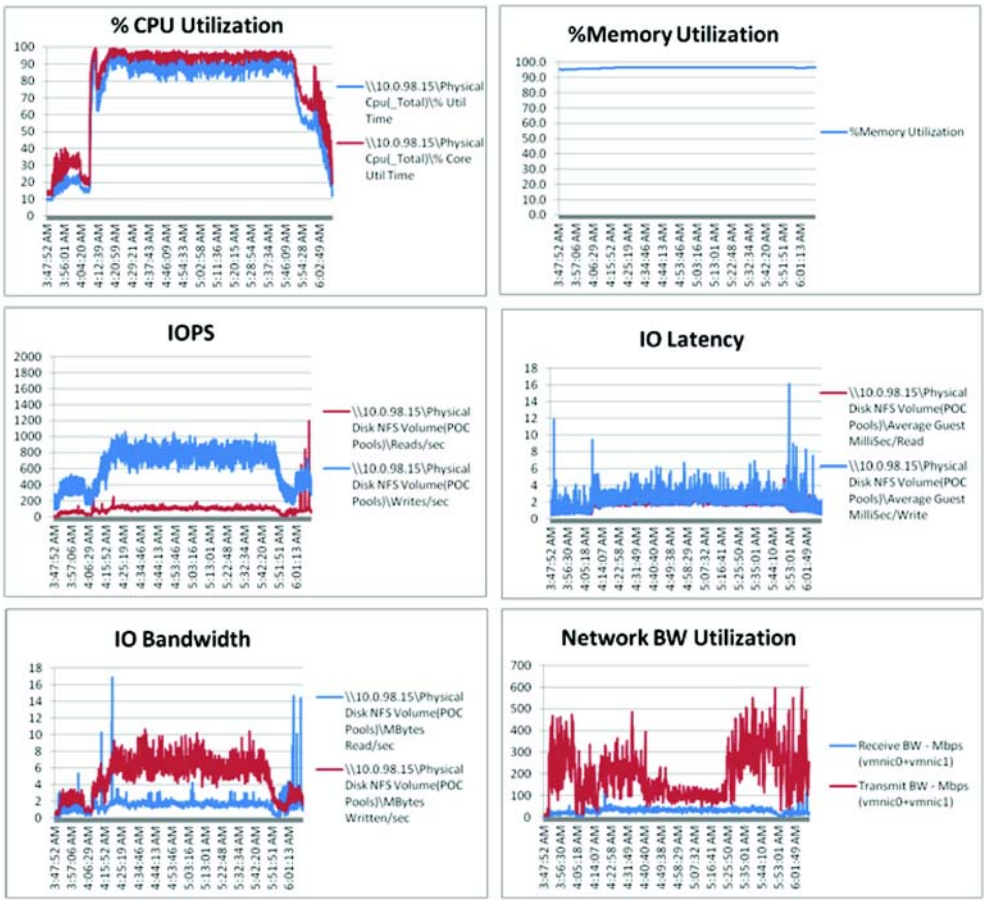
- NAS NFS
- NetApp FAS 3170 with PAM 2 module (512G of cache)

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login, Workload and Logout phases

Performance Charts

Figure 40 Performance Charts for View5/ESXi5/PCoIP/B230M2/NetApp Profile



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 11 Response Times for View5/ESXi5/PCoIP/B230M2/NetApp Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	2.2s
Outlook	5s	3.5s
Word	5s	2.3s
Excel	5s	2.6s
PowerPoint	5s	1.5s

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Internet Explorer	5s	2.5s
Acrobat	5s	1.4s

Scale and Performance Baseline for VMware View without Antivirus

The objective of the scale and performance testing with this profile is to provide baseline guidance in terms of virtual desktop density supported on a UCS server without antivirus. Results indicate that a density of ~110 virtual desktops can be supported on a UCS B250 M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are CPU bound for this profile.

Test Profile

Desktop Virtualization

- VMware View 4.6
- Connection Protocol – PCoIP
- Linked Clones

Hypervisor

VMware ESXi 4.1U1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 20G disk configured per desktop
- 1 vCPU per desktop
- Persistent desktop

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon X5680 processors @ 3.33 GHz
- 192 RAM (48 x 4G DIMMS @1333MHz)
- UCS M81KR Virtual Interface Card/PCIe/2-port 10Gb

Workload Profile: Cisco Knowledge Worker+ (v2.5)

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in deskphone mode

Storage

- VSPEX (EMC VNX 5500) - Fibre Channel

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Resxtop with a polling interval of 5s
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login, Workload & Logout phases

Summary of Test Results

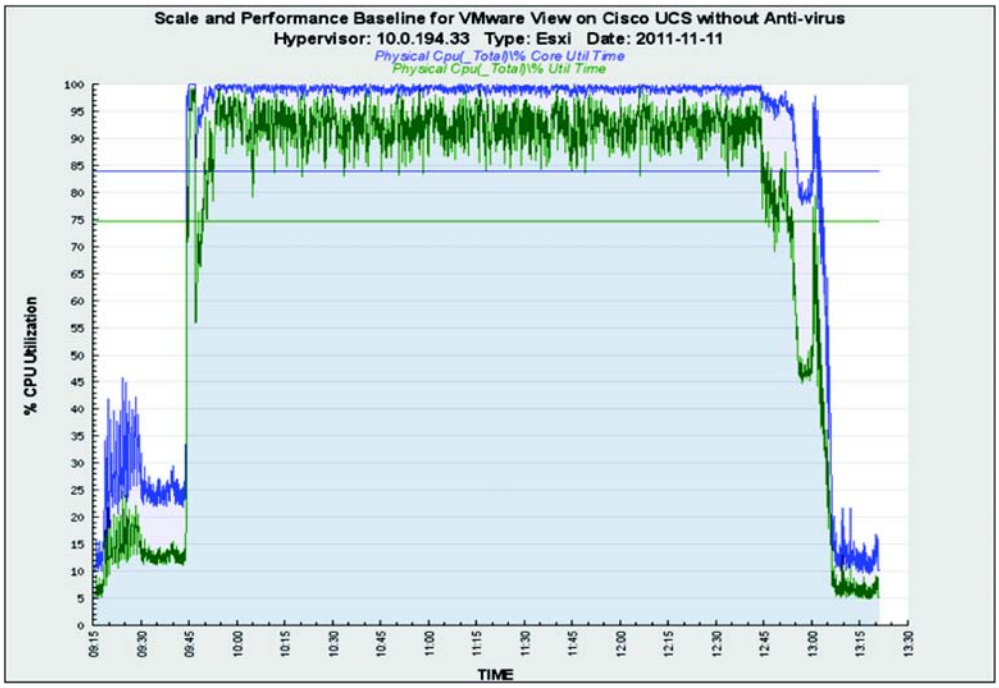
For the deployment profile detailed above, 113 VMs can be supported on a Cisco UCS B250 M2 with the following performance metrics.

Server Metrics:

- Average CPU Utilization = ~92% (Steady state)
- Average Memory Utilization = ~93%
- Application Response times – Success Criteria met

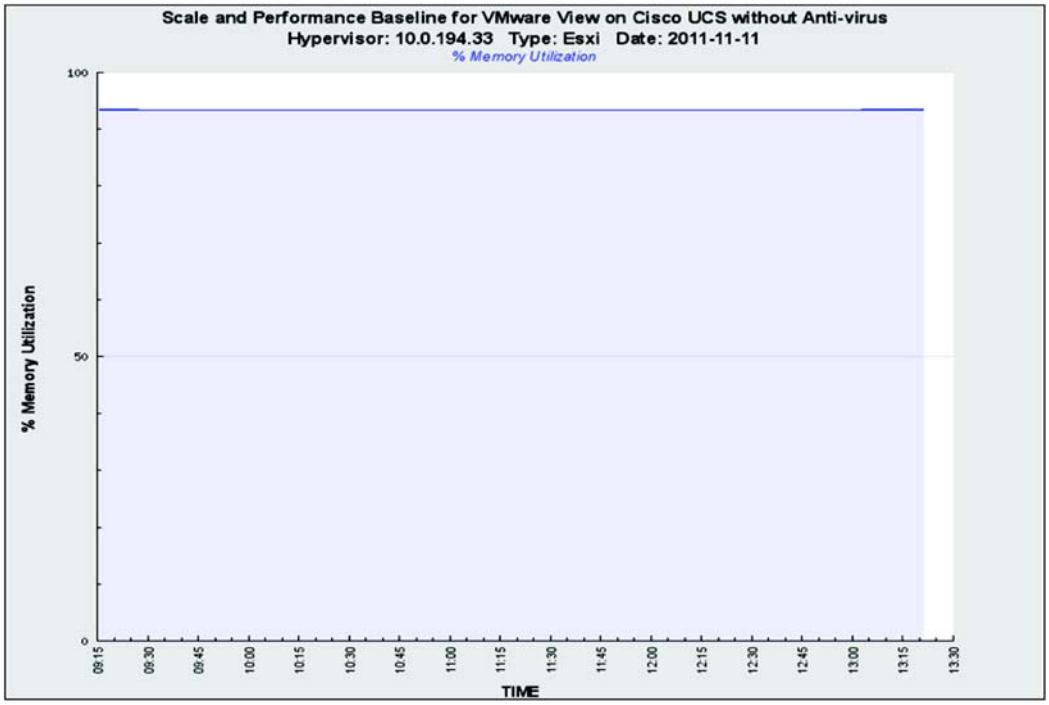
CPU Utilization

Figure 41 CPU Utilization Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)



Memory Utilization

Figure 42 Memory Utilization Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)



IO Statistics

Figure 43 IOPS Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)

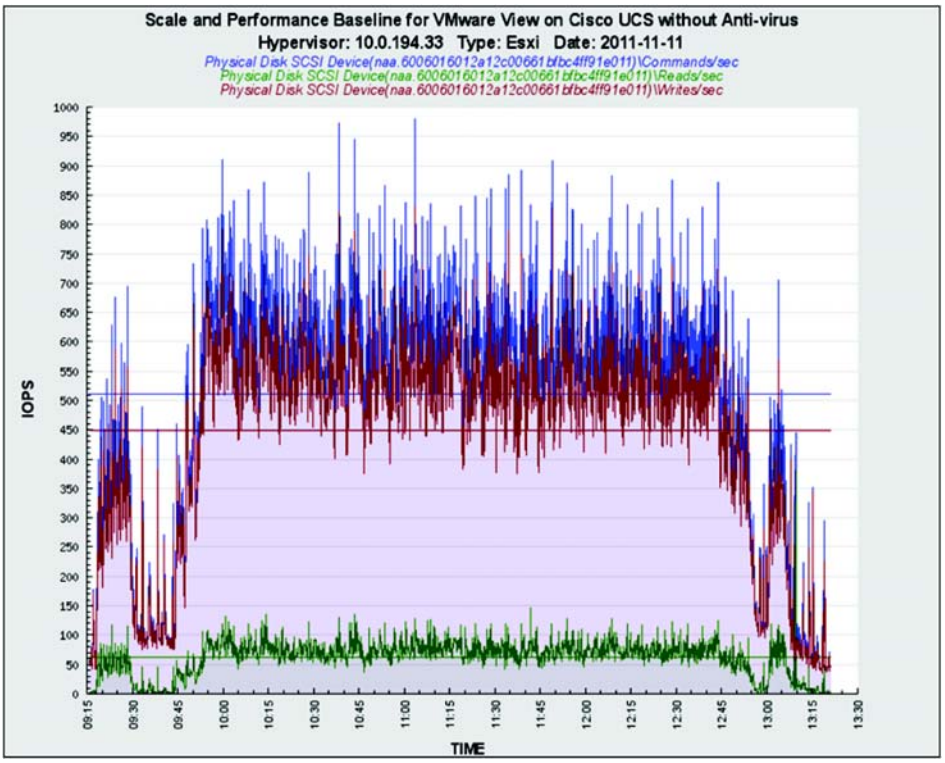


Figure 44 IO Latency Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)

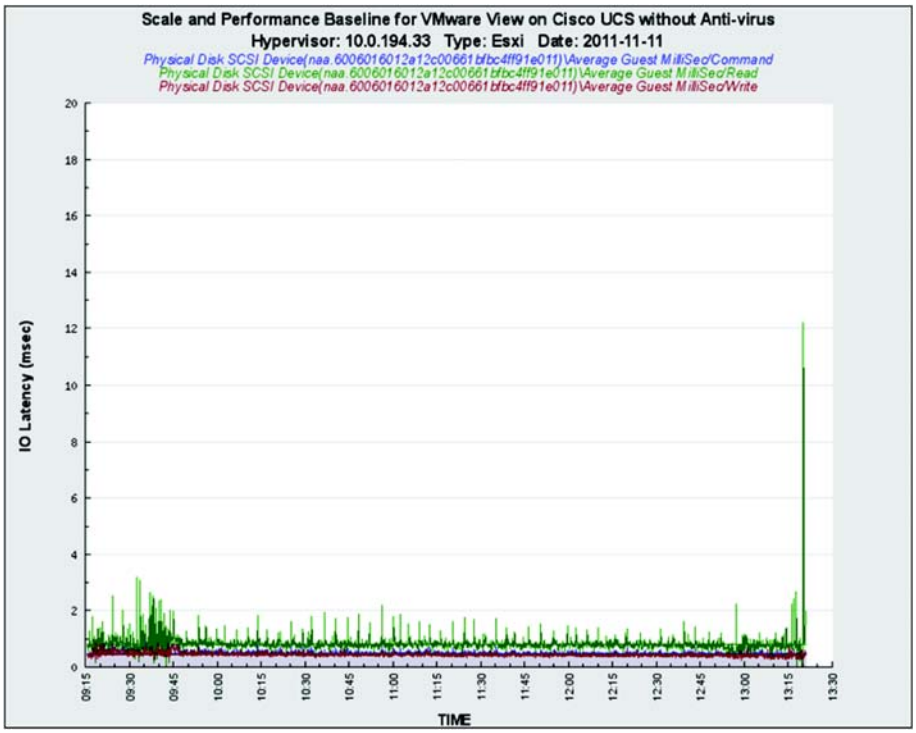
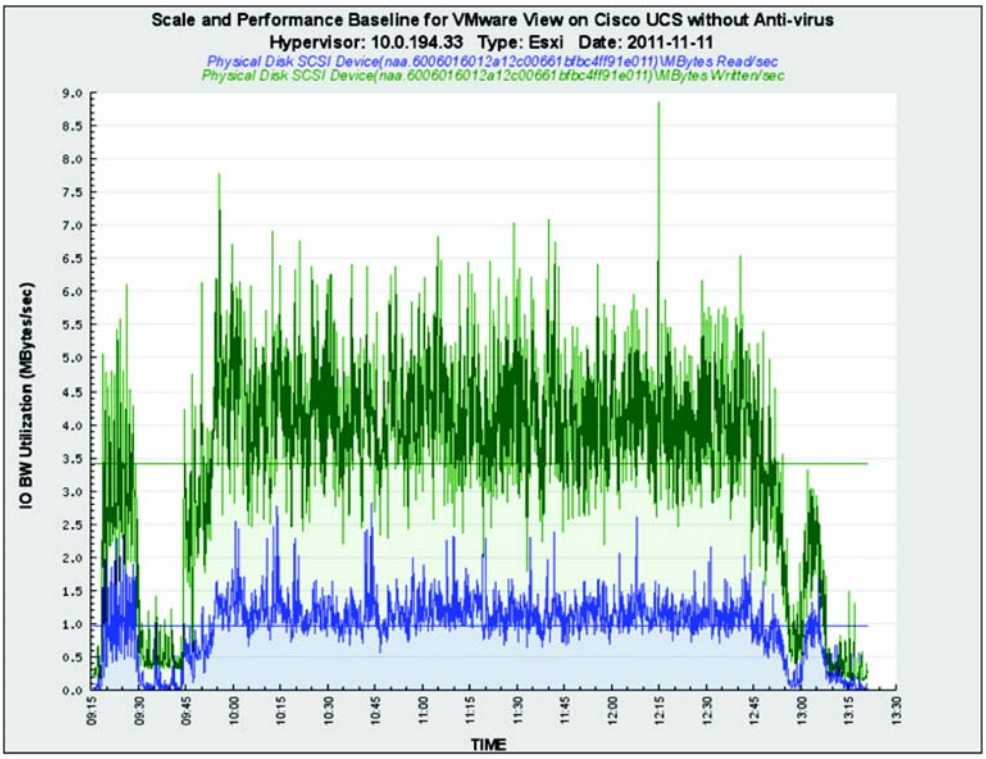
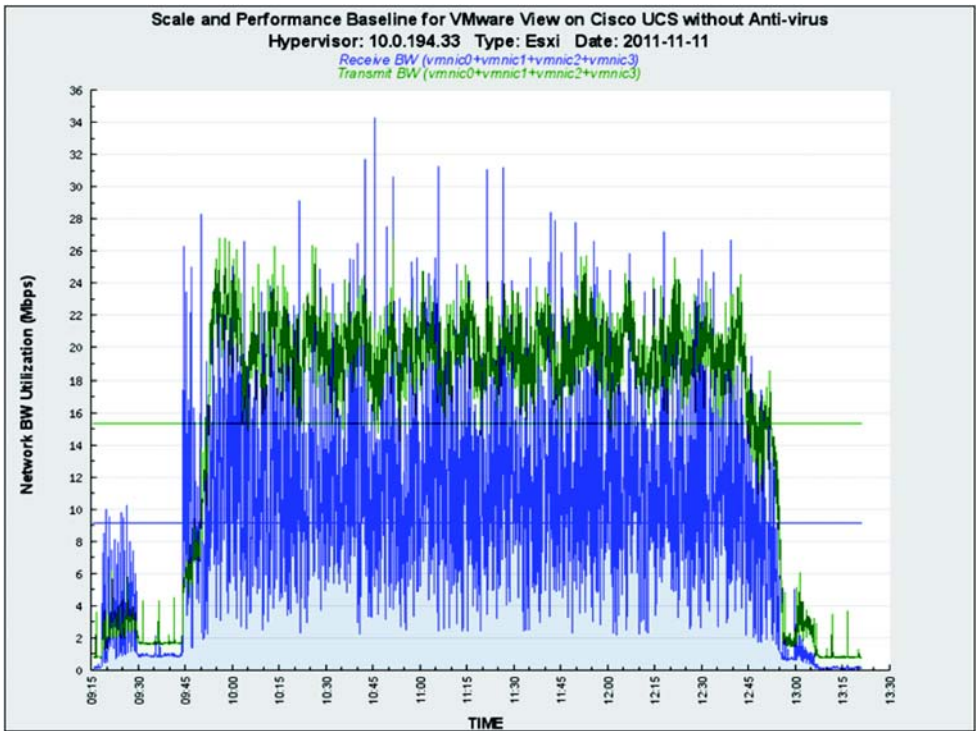


Figure 45 IO BW Utilization Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)



Network Bandwidth Usage

Figure 46 Network BW Utilization Chart for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)



Application Response Times

The response times for this profile were well within the success criteria defined at the beginning of this document.

Table 12 Response Times for Baseline Profile (View4.6/ESXi4.1/PCoIP/B250M2/EMC)

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	1.9s
Outlook	5s	4.0s
Word	5s	2.4s
Excel	5s	2.6s
PowerPoint	5s	1.5s
Internet Explorer	5s	2.4s
Acrobat	5s	1.5s

Network Characterization

This section focuses on deploying desktop virtualization users at branch sites across an Enterprise WAN and the validation data needed to guide your WAN capacity planning. The following three aspects will be covered here:

- High level summary of deployment profiles tested
- Validation methodology
- Detailed test results

Summary of Results

In this section, a high level summary of the areas characterized from a WAN capacity planning perspective across the end-to-end Cisco Virtual Workspace Solution are provided in the Table 13 below.

Table 13 WAN Capacity Planning

Desktop Virtualization	Workload Profile	HVD Profile	Storage	UCS Server
Objective: Understanding the bandwidth (BW) characteristics of a Cisco KW+ workload				
View 4.5 on ESXi 4.1 (PCoIP)	Cisco Knowledge Worker+	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)

Desktop Virtualization	Workload Profile	HVD Profile	Storage	UCS Server
View 4.5 on ESXi 4.1 (RDP)	Cisco Knowledge Worker+	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)
Objective: Understanding the bandwidth characteristics of a video-only workload				
View 4.5 on ESXi 4.1 (PCoIP)	Video-only	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)
View 4.5 on ESXi 4.1 (RDP)	Video-only	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)
Objective: Impact of display protocol adaptiveness on server/compute performance at scale				
View 4.5 on ESXi 4.1 (PCoIP)	Cisco Knowledge Worker+	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)
View 4.5 on ESXi 4.1 (RDP)	Cisco Knowledge Worker+	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)
Objective: Impact of WAAS Optimization on WAN deployments with View RDP				
View 4.5 on ESXi 4.1 (RDP)	Cisco Knowledge Worker+	Win 7 32b (1.5G, 20G, 1vCPU)	T1 with 80ms of latency	B200 M2 (2 x 6 core X5680 @3.33 GHz with 96G of memory)

Validation Methodology

The methodology used for characterizing Cisco Virtual Workspace deployments across the WAN is similar to the validation methodology outline in Single Server Scale and Performance section of this document. However, since the objective is not to determine the max density at the server level, the success criteria does not look at the CPU or memory utilization except in the case of two tests documented below. All testing is done across the end-to-end Cisco Virtual Workspace system and in this case across a WAN link to branch site. Workload profile used in all cases is the Cisco KW+ profile – however there is more emphasis paced on subjective user experience in addition to application response timers.

Detailed Test Results

A detailed analysis of the test results and the associated profile and objectives are provided in this section.

Bandwidth Characteristics of a DV workload – Cisco KW+ workload

In this section, we explore the bandwidth characteristics of a typical DV workload using Cisco KW+ as an example. In a desktop virtualization environment, the workload profile used is a critical component of performance characterization, including bandwidth characterization, which is the focus here. As the bandwidth characteristics can vary with the workload, in an actual deployment, it is important to do a similar assessment using a workload that closely matches the customer's environment. The workload used should be representative of their user base, not only in terms of applications but also with respect to usage patterns. Having said that, the Cisco KW+ workload is very representative of a typical knowledge worker, both in terms of the applications (Microsoft Office Applications, Internet Explorer, Adobe Acrobat) and in terms of the operations within these applications so the data here should provide a good basis for sizing WAN links in any Cisco Virtual Workspace deployment.

The bandwidth data provided in this section are as follows:

- The peak bandwidth for a given workload and user with unrestricted bandwidth. This testing is done across a T1 link with one user at the branch site across a Cisco Virtual Workspace network with the HVD hosted on a UCS blade in the data center. A delay of ~80ms is injected on all traffic across the WAN link and it represents the typical latency seen from East Coast to West Coast in the US. Since all of the T1 bandwidth is available for a single user, the bandwidth should be sufficient to handle the average BW utilization for Knowledge worker especially but may not be enough to handle peaks in the workload – see next bullet point that addresses this.
- Differences in the peak bandwidth utilization seen with the workload when the same user is in a campus network with high speed links (>T1) with enough BW to handle the peaks.
- Application level break down of BW consumption, including BW required to login and logout of an HVD. This provides not only relative BW consumption data between user applications such as Word, Excel but also as it related to DV specific activities such as HVD login and logout. In addition, the data also provides information on actions within an application and its impact to bandwidth usage.
- Minimum bandwidth required for the given workload so that good UE is still maintained. This bandwidth can be the basis for any WAN sizing in an environment with similar workload.

Test Environment and Setup

- View 4.5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory
    - Display protocol: PCoIP, RDP
    - Display Session Characteristics:
  - Screen Resolution: 1366x768 (Large Window)
  - Color Depth: 16bit
  - Windows optimized for Best Performance (All Options checked off)
- Workload Profile: Cisco Knowledge Worker+ profile with optimized antivirus solution from a leading vendor
- A single HVD was used for this test
- Server Profile: UCS B200 M2 with 96G of RAM – server was running at minimal loads during this test

- For this test, a single HVD from a branch site across a T1 WAN link was used. Delay of 80ms was injected but no jitter

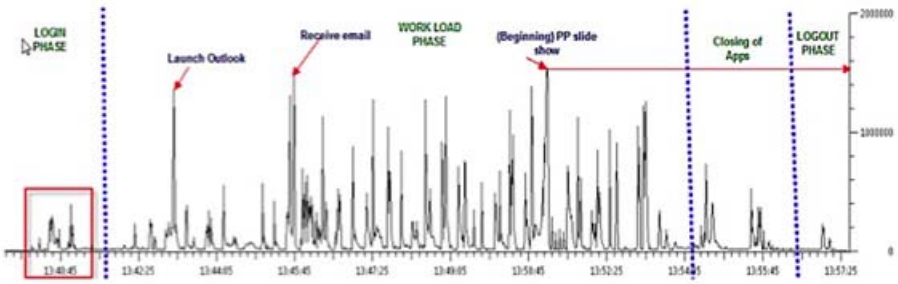
User Experience/Application Response Times

For this test, the user experience was observed over multiple iterations of a given test run while an automated workload was executed across the display session. The session experience was also recorded on WebEx for additional review and analysis. In this particular case with only one user, the subjective measurements are a better gauge of true user experience as it captures all aspects of the session experience while a test tool may only capture response times for certain activities. In addition, as the testing progressed, it also became obvious that certain activities within the applications in the workload are more susceptible to issues that impact user experience and therefore careful attention was paid to these areas when assessing the session experience. Examples of this include the viewing a PowerPoint in Slide Show mode and the composing of an email in Outlook. When bandwidth restriction starts to impact UE, the information on a PowerPoint slide can get presented in blocks while in Outlook, the message being typed can get displayed in chunks as opposed to a smooth flow of words when there are no user experience issues. In summary, the results of this test are based on subjective user experience but in this particular case, the bandwidth characterization data should be more reliable and accurate because all aspects of the session experience is being observed.

Summary of Test Results

Bandwidth – Peak and Average

Figure 47 Bandwidth Utilization for a PCoIP session with Cisco KW+ workload



Display
Display filter: ip.addr == 10.1.3.10

Traffic	Captured	Displayed	Marked
Packets	40089	39247	0
Between first and last packet	1125.443 sec	1056.956 sec	
Avg. packets/sec	35.621	37.132	
Avg. packet size	364.652 bytes	370.954 bytes	
Bytes	14618526	14550818	
Avg. bytes/sec	12989.132	13774.296	
Avg. MBit/sec	0.104	0.110	

255619

The above figure shows the bandwidth during a single iteration of the automated Cisco KW+ workload where the workload represents a user's activities during that time frame. The data is from a single DV session with no other traffic on the link other than minimal control traffic and the graph above is filtered view to show just the DV session traffic. The information also shows the bandwidth utilization when the user first logs into a DV session and when the user logs out. The peak bandwidth utilized in each phase is summarized in the table below. Note that the workload peaks are the highest, hitting T1 speeds, followed by the login phase. Logout phase seems to have the least BW impact among the three phases.

Since a T1 WAN link was used for these tests, the peak bandwidth associated with the remote displaying of any event in the workload cannot be higher than a T1. Therefore depending on the display protocol and the bandwidth requirements of this workload, the peaks may not be the true peak for the workload if the display protocol already adapted due to the T1 limit. The same tests repeated from a campus location with 100Mbps+ bandwidth will confirm whether this is the true peak for the workload or post-adjustment peak – see below. suffer during the workload peak though it may have been limited by the T1 link.

Table 14 Peak Bandwidth for a single DV session using PCoIP and a Cisco KW+ workload across a T1

Branch	Peak BW Run #1	Peak BW Run #2	Peak BW Run #3	Peak Bandwidth Usage
Login	400 kbps	500 kbps	400 kbps	433 kbps
Workload	Full T1	Full T1	Full T1	Full T1
Logout	290 kbps	350 kbps	350 kbps	330 kbps

The above data for a single user using a given workload can now be used in conjunction with the minimum BW data to define the bandwidth range that provides good UE – this data is key to the sizing the WAN link for a branch Cisco Virtual Workspace deployment.

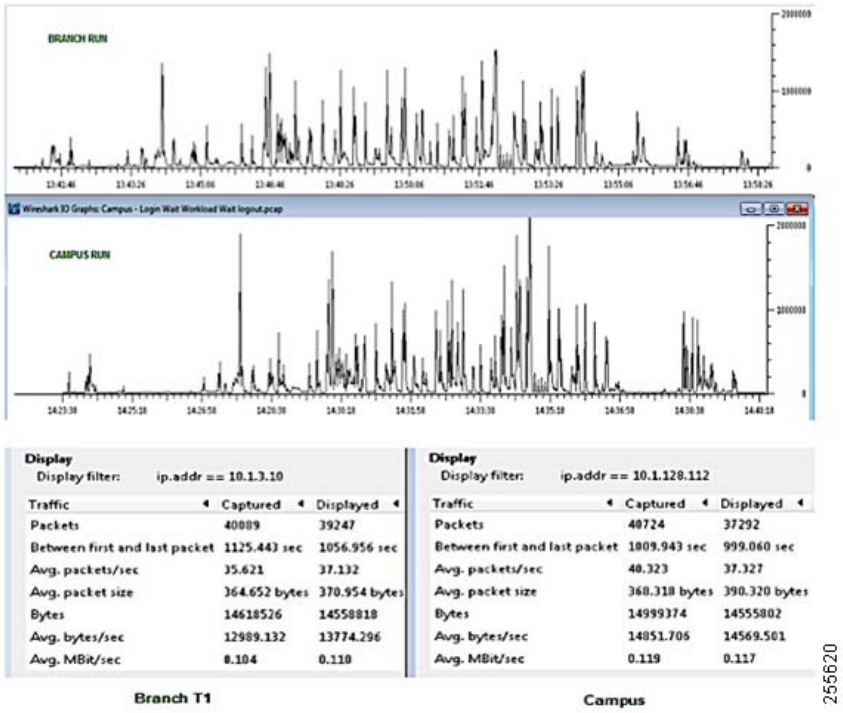
The above figure also shows the workload further detailed in terms of the applications and activities within the workload. This shows both the absolute and relative BW impact that a given application or action within the application can have when it is remotely displayed to the user. Note that peak bandwidth during this workload is seen from PowerPoint in Slide Show mode and Outlook. It is also important to note that from a user experience perspective, typing of an email though it uses less bandwidth is very susceptible to UE issues when there is bandwidth congestion. On the other hand, UE was not impacted when OL was sending/receiving though it is the second biggest consumer of bandwidth. User Experience did get affected during the start of the slide show in PowerPoint.

Similar to PCoIP, the peak bandwidth for RDP, measured at the branch is summarized in the table below.

Table 15 Peak BW for a single DV session using RDP and a Cisco KW+ workload across a T1

Branch	Peak Bandwidth Usage
Logn	800 kbps
Workload	Full T1
Logout	600 kbps

Figure 48 Branch vs. Campus View of Bandwidth Utilization for PCoIP



RDP

Figure 49 Branch vs. Campus View of Bandwidth Utilization for RDP – Campus

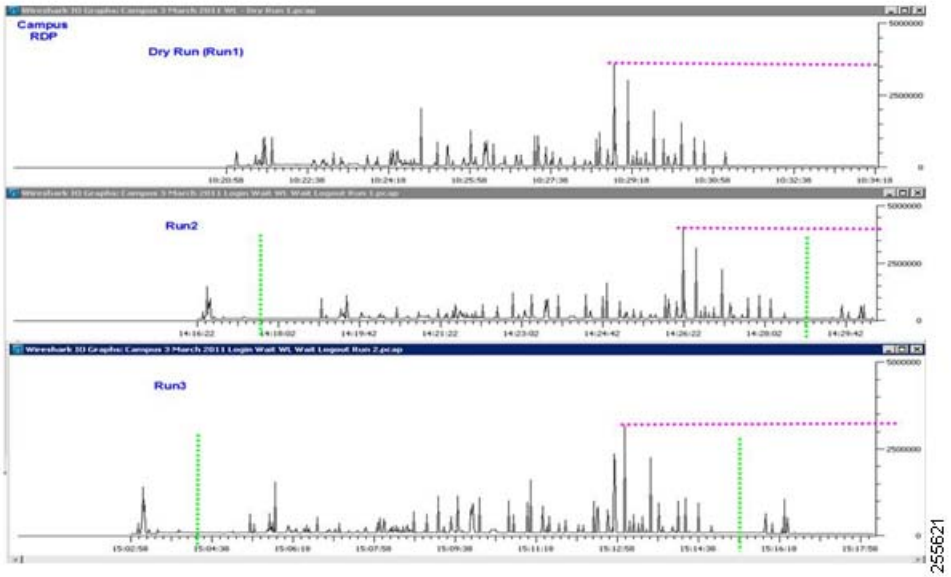
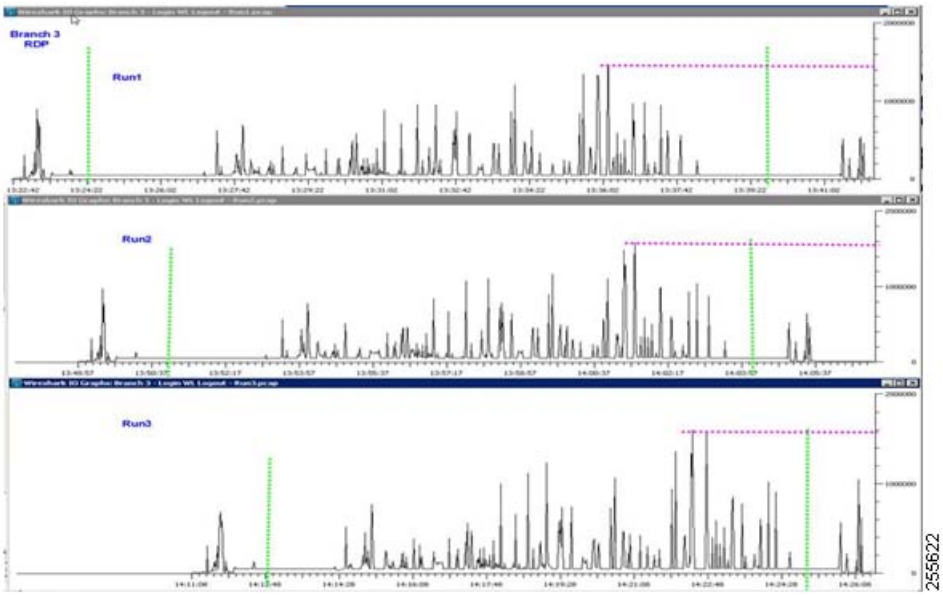


Figure 50 Branch vs. Campus View of Bandwidth Utilization for RDP – Branch



The three above figures shows the peak bandwidth usage for both PCoIP and RDP when going from a T1 with a single user (and delay of 80ms) to a Campus with 100Mbps+ bandwidth (and no delay). The graphs clearly show that the peak bandwidth seen for the same workload is actually higher though user experience may not have suffered at T1 speeds. So if you are sizing a WAN link to accommodate the peaks or to an X% of that peak, it is important to determine the peak bandwidth in an environment where there is enough bandwidth to handle the peaks. Note that if the sizing were based on the average bandwidth utilized by the workload, this would not be a concern with only a single user on a T1.

Based on the above, the data from branch testing can be updated for PCoIP to reflect the true peak BW during the workload phase as follows:

Table 16 Peak Bandwidth for a single DV session using PCoIP and a Cisco KW+ workload

	Branch - Peak BW	Campus - Peak BW
Logn	567 kbps	Same
Workload	967 kbps	~1.3 Mbps
Logout	293 kbps	Same

Similarly, the peak bandwidth seen for branch and campus can be summarized as follows.

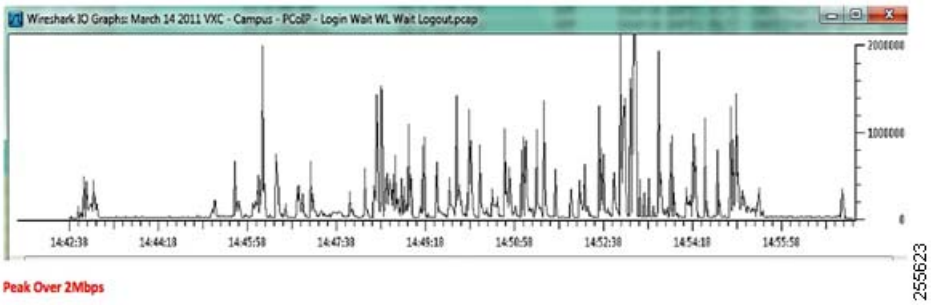
Table 17 Peak Bandwidth for a single DV session using RDP and a Cisco KW+ workload

	Branch - Peak BW	Campus - Peak BW
Logn	567 kbps	Same
Workload	967 kbps	~1.3 Mbps
Logout	293 kbps	Same

Branch Versus Campus

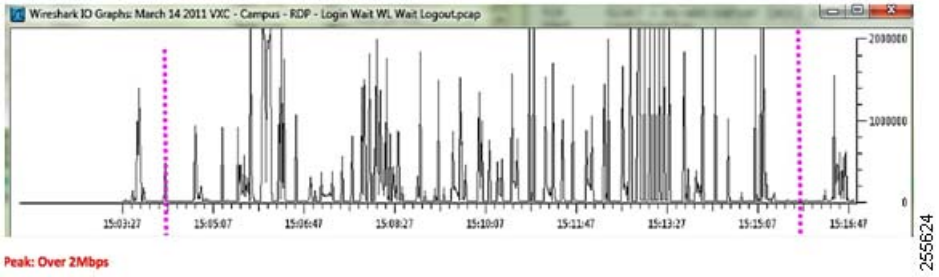
In the figure below, the data in the table above is further confirmed by using a Cisco VXC endpoint in campus to initiate a PCoIP session and measuring the bandwidth used by the single session when activities similar to the automated workload are done manually such as PowerPoint Slide Shown and send/receiving of email in Outlook.

Figure 51 PCoIP Maximum BW for Cisco VXC client running KW+ workload



Similarly, the data above is confirmed for RDP using a Cisco VXC endpoint.

Figure 52 RDP Maximum BW for Cisco VXC client running KW+ workload



Minimum Bandwidth

To determine the minimum bandwidth necessary to provide good user experience with this workload, the available bandwidth on the T1 is reduced until the user experience suffers. In this case, removing the timeslots from the channelized T1 link was used to reduce the available bandwidth. The automated workload is then run and when the user experience starts to become unacceptable, the bandwidth on the T1 just before this point is assumed to be the minimum bandwidth.

Figure 53 PCoIP Minimum BW for Cisco VXC client running KW+ workload

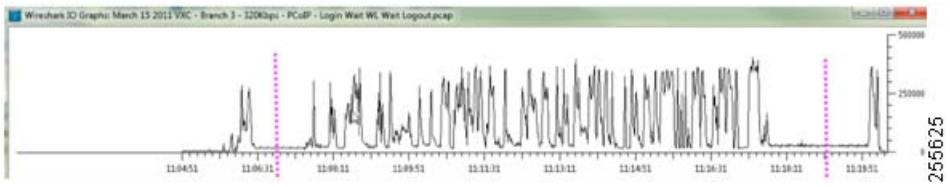
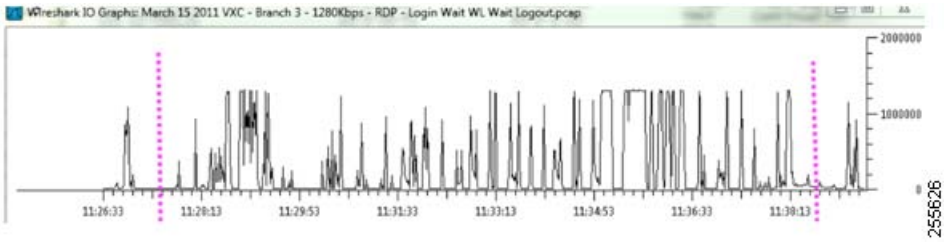


Figure 54 RDP Minimum BW for Cisco VXC client running KW+ workload



Using the above methodology, the minimum bandwidth for good user experience when the display protocol is PCoIP is 320kbps and for RDP, it is 1280kbps for the same workload.

Bandwidth Characteristics of a Video Only DV workload

This section focuses on the bandwidth characteristics of a video only DV workload to understand the impact that a short video clip can have on the bandwidth requirements of a branch site. For these tests, a one-minute flash video clip was used across a WAN link (T1 in this case) and the user experience is observed with and without congestion. As in the previous case, bandwidth available for the DV session is reduced to create the congestion.

Test Environment and Setup

- View 4.5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory
  - Display protocol: PCoIP, RDP
  - Display Session Characteristics:
    - Screen Resolution: 1366x768 (Large Window)
    - Color Depth: 16bit
  - Windows optimized for Best Performance (All Options checked off)
- Workload Profile: Video only – 1 min. Flash video clip, Standard Definition, 640x360
- A single HVD is used for this test
- Server Profile: UCS B200 M2 with 96G of RAM – server was running at minimal loads during this test
- For this test, a single HVD from a branch site, across a T1 WAN link was used. Delay of 80ms was injected but no jitter

Summary of Test Results

Bandwidth – Peak and Average

The four figures below show the bandwidth utilization of a 1min video clip without congestion for both PCoIP and RDP. Note that that the average and peak utilization of this video workload is the full available T1 bandwidth. The user experience, both video and audio quality was acceptable for this test. However, video will use as much of the available bandwidth as it needs and has stringent loss, jitter and latency requirements. It can also starve out other application traffic or its quality can be impacted by other traffic. In any case, if video is part of the branch Cisco Virtual Workspace deployment, the overall design approach cannot be based on determining a minimum BW and then sizing based on that as in the

previous Cisco KW+ workload case. It needs to factor in the types of video formats that needs to be supported, whether the video is carried in band or outside of the display session, specific technologies available in the context of desktop virtualization and associated optimization technologies, ability to implement network level QoS policies as well content caching solutions that are outside the framework of desktop virtualization but that can integrate with user requests from within the display protocol session. In short, the exercise here merely confirms that deployment and design of video in branch Cisco Virtual Workspace deployments requires a different approach due to its bandwidth characteristics. Also, if the video traffic must traverse the WAN link to the branch site, ability to provide network level QoS is a must in any design options being considered.

PCoIP

Figure 55 Bandwidth Utilization for a PCoIP session with Video-only workload

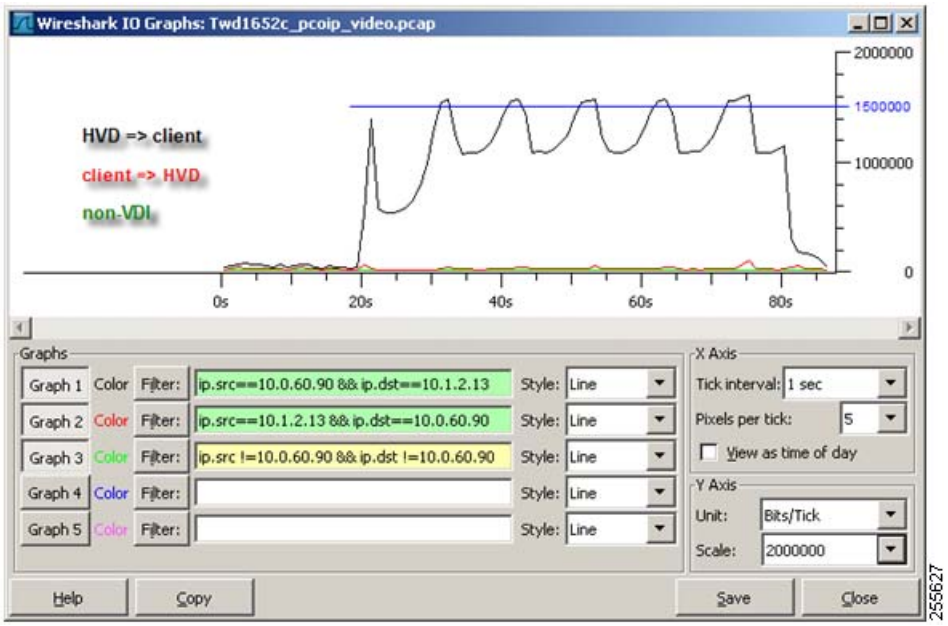
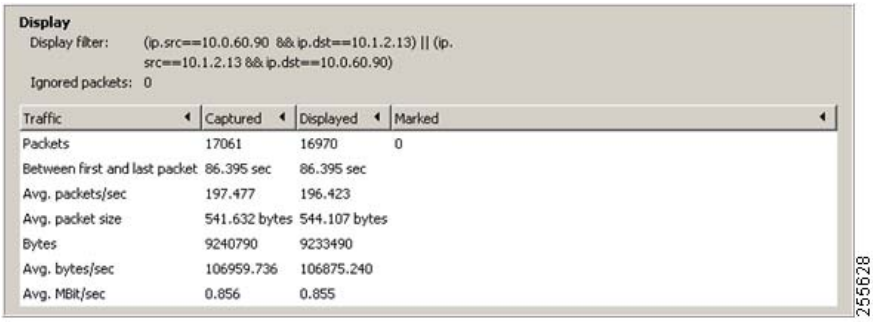


Figure 56 Average Bandwidth Utilization Stats for a PCoIP session with Video-only workload



Display

Display filter: (ip.src==10.0.60.90 && ip.dst==10.1.2.13) || (ip.src==10.1.2.13 && ip.dst==10.0.60.90)

Ignored packets: 0

Traffic	Captured	Displayed	Marked
Packets	17061	16970	0
Between first and last packet	86.395 sec	86.395 sec	
Avg. packets/sec	197.477	196.423	
Avg. packet size	541.632 bytes	544.107 bytes	
Bytes	9240790	9233490	
Avg. bytes/sec	106959.736	106875.240	
Avg. MBit/sec	0.856	0.855	

256628

RDP

Figure 57 Bandwidth Utilization for a RDP session with Video-only workload

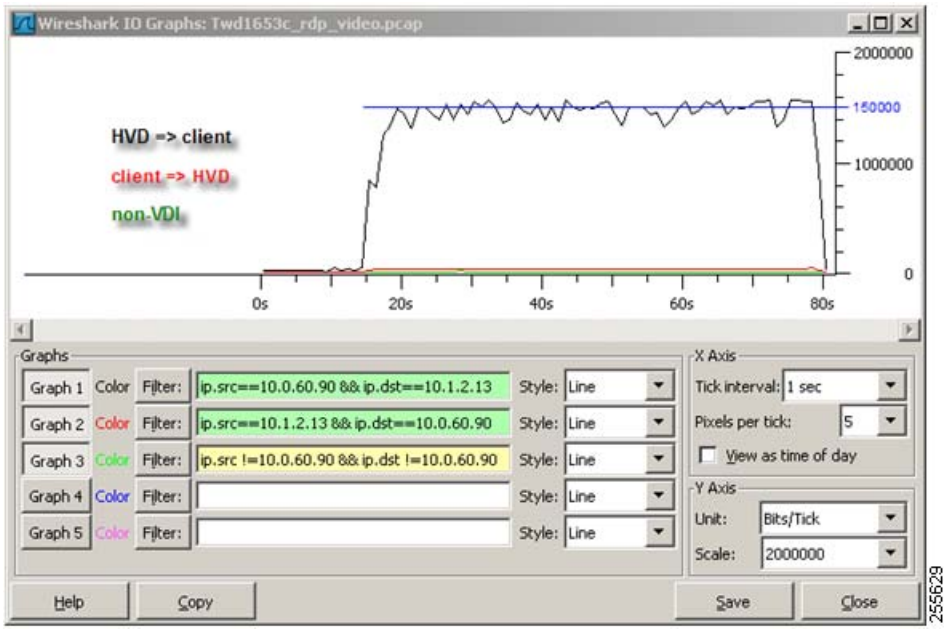
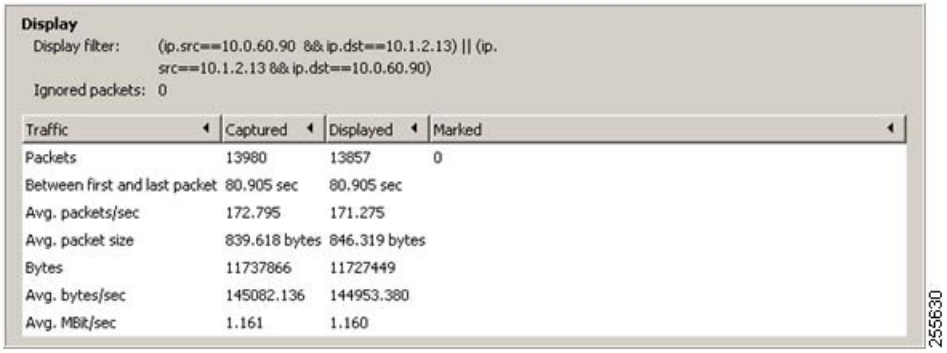


Figure 58 Average Bandwidth Utilization Stats for a RDP session with Video-only workload



Minimum Bandwidth

The four figures below show the bandwidth utilization of a 1min video clip with congestion using PCOIP and RDP. Note that that the average and peak utilization during the workload phase continues to take up the full available bandwidth, which in this case was reduced to 768kbps for PCoIP and 1024kbps for RDP. However, the user experience, both video and audio quality suffered at these rates. For PCoIP, audio was clear but not in sync with video. Video was choppy at times but marginally acceptable. Even with loss of audio/video sync, content was understandable and video was watchable.

Video was choppy, difficult to understand and audio was out-of-sync with the video. For RDP with 1024kbs worth of BW, video was choppy throughout, audio not in sync and so it fared worse and the experience is marginally acceptable.

Based on this, a T1 worth of bandwidth is necessary even with a short 1 min, Standard Definition (640x360) clip though it could go down as low as 768kbps (PCoIP) and 1024kbps (RDP) but will suffer from the quality issues described above.

PCoIP

Figure 59 Minimum BW for a PCoIP session with Video-only workload

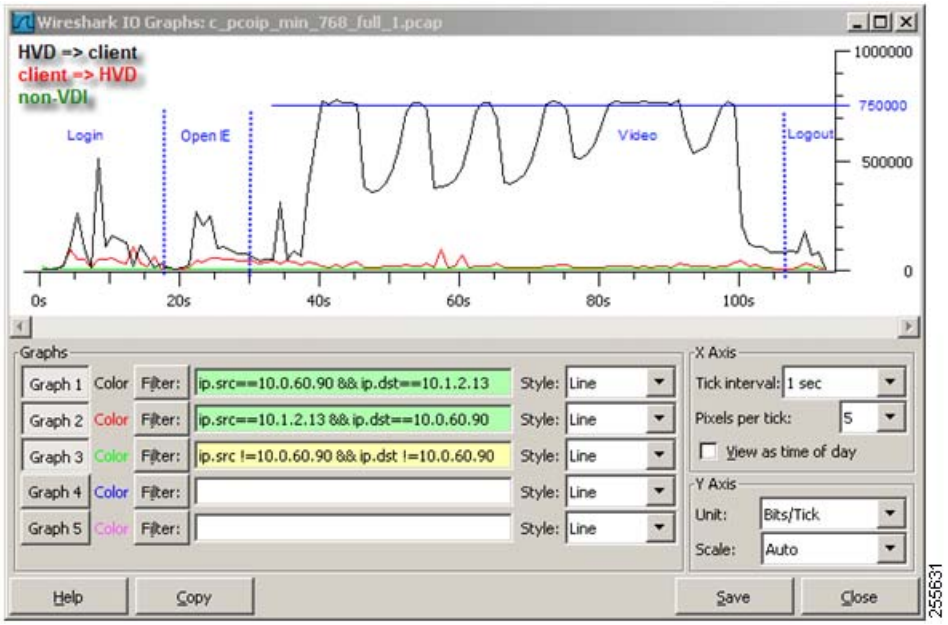
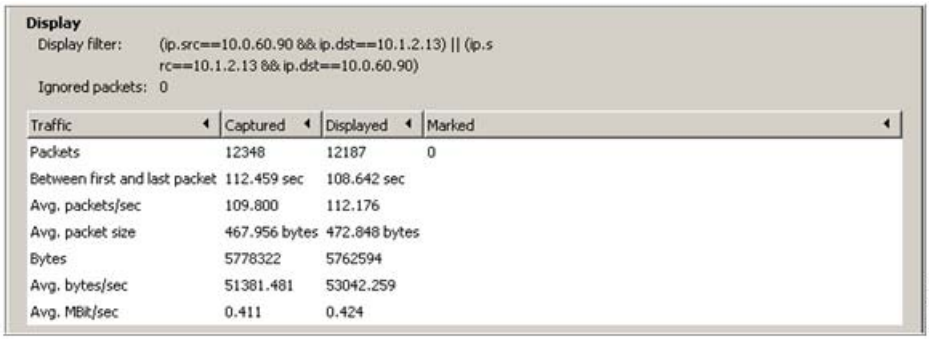


Figure 60 Minimum BW Statistics for a PCoIP session with Video-only workload



RDP

Figure 61 Minimum BW for a RDP session with Video-only workload

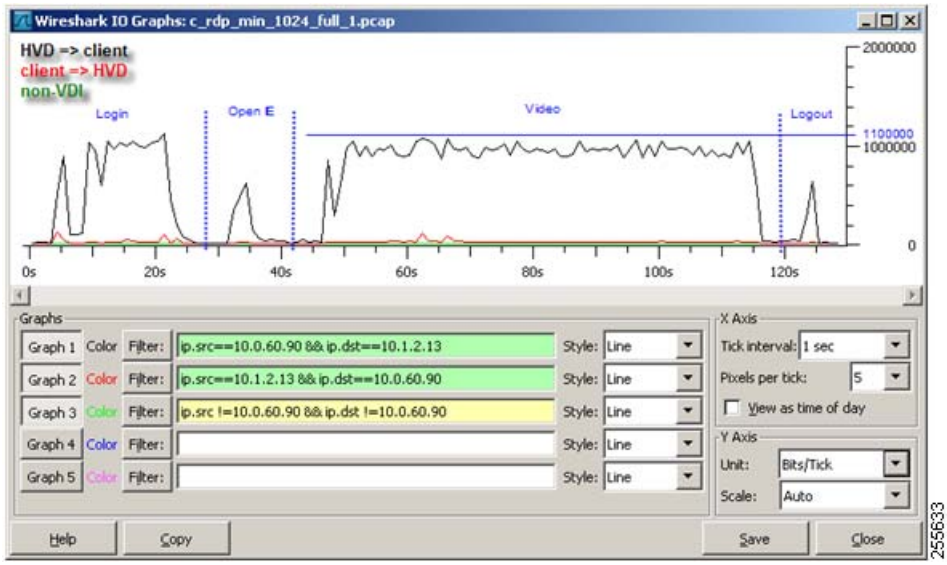
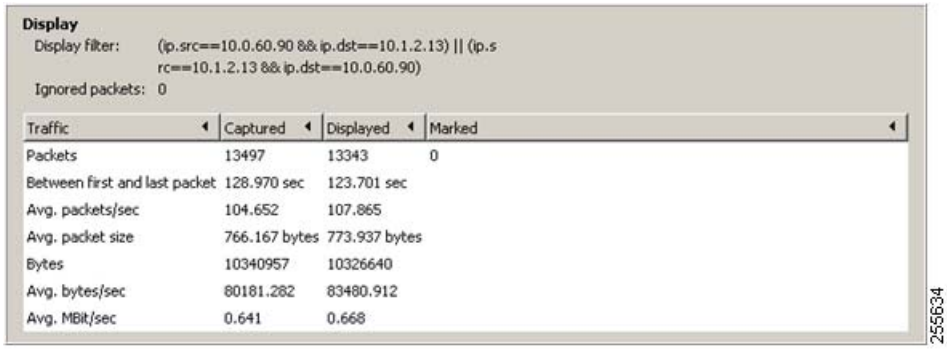


Figure 62 Minimum BW Statistics for a RDP session with Video-only workload



Impact of Protocol Adaptiveness on Server/Compute Performance

For large branch based Cisco Virtual Workspace deployments, as congestion occurs and display protocols start adapting, it is important to understand whether the adaptive nature of the display protocols has any impact on the server hosting the virtual desktops. Server

scale and performance benchmarking typically does not factor this in and is done without any bandwidth constraints. Therefore, the objective here is to determine the impact, to a server running at full load where all users are in branch sites, the impact of congestion and display protocol adaptiveness.

Test Environment and Setup

- View 4.5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory
  - Display protocol: PCoIP, RDP
  - Display Session Characteristics:

    - Screen Resolution: 1350 x 686

    - Color Depth: 16bit
  - Windows optimized for Best Performance (All Options checked off)
- Workload Profile: Cisco Knowledge Worker+ profile with optimized antivirus solution from a leading vendor
- Server Profile: UCS B250 M2 with 192G of RAM – server was scaled to maximum capacity and running at 90% CPU utilization.
- For this test, all HVDs hosted on the UCS server were accessed from branch sites, across T3 WAN links. Delay of 80ms was injected but no jitter

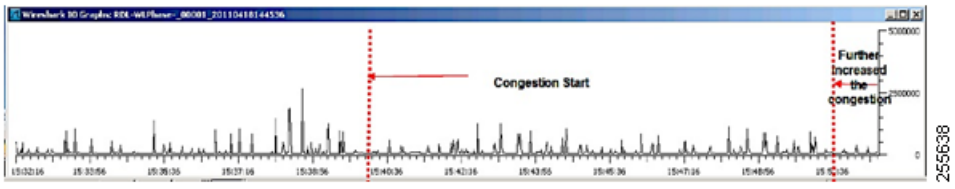
Summary of Test Results - PCoIP

The graph below shows the CPU utilization on a Cisco UCS B250 M2 server hosting 80 HVDs where all users are in branch networks across the Cisco Virtual Workspace network. With this number of hosted virtual desktops, the CPU utilization is at a steady state of ~90% utilization and at steady state, congestion is introduced on the WAN links using a traffic generator. The results show that there is no impact on the server performance as the sessions adapt down to use less bandwidth – see graph below. The user experience in terms of application response times were also measured for each application across all 80 sessions and were well within the acceptable range.

Figure 63 Bandwidth Utilization on UCS B250 server at scale before and after congestion with PCoIP

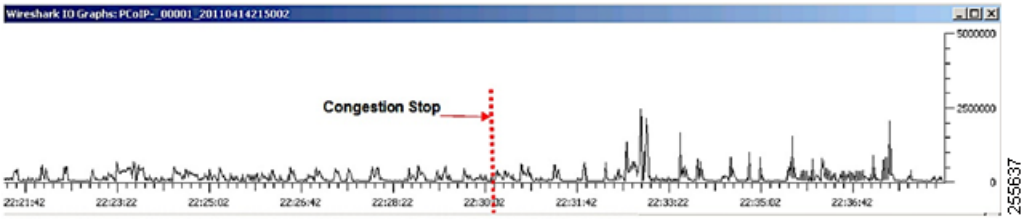
This graph shows the bandwidth utilization measured for one session on the UCS server. Until congestion starts, the PCoIP session was using the peak bandwidth measured in bandwidth characterization section earlier and adapts down as congestion increases.

Figure 64 Bandwidth Utilization before and after congestion start with PCoIP – single session



It is also worth noting that as the congestion traffic is reduced, PCoIP session quickly ramps up its bandwidth usage and is already at 2.5Mbps within a few seconds.

Figure 65 Bandwidth Utilization before and after congestion stops with PCoIP



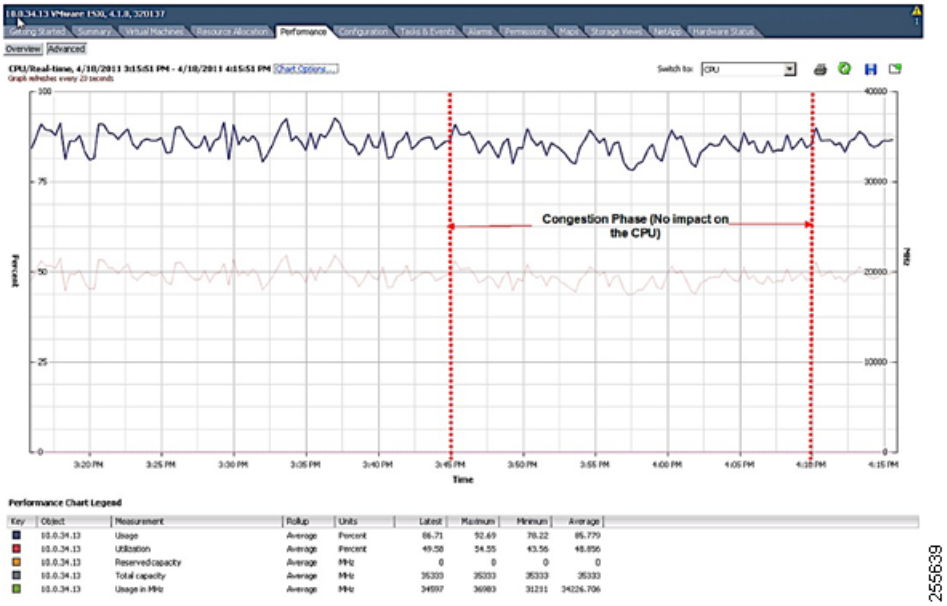
Summary of Test Results

RDP

Similar to PCoIP, a UCS B250 M2 blade is deployed with 85VMs running at 85% CPU utilization and the impact to its performance is measured as congestion starts. The graph below shows the CPU utilization of the server running 85 HVDs where all users are in branch networks across the Cisco Virtual Workspace network. As congestion starts, the graphs show that there is no impact on the server

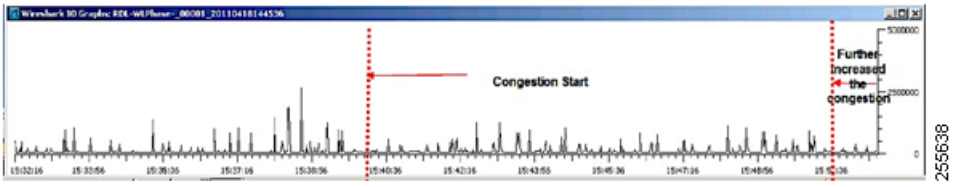
performance as the RDP sessions adapt down to use less bandwidth. The user experience in terms of application response times is also measured for each application across all 80 sessions and was well within the acceptable range.

Figure 66 Bandwidth Utilization on UCS B250 server at scale before and after congestion with RDP



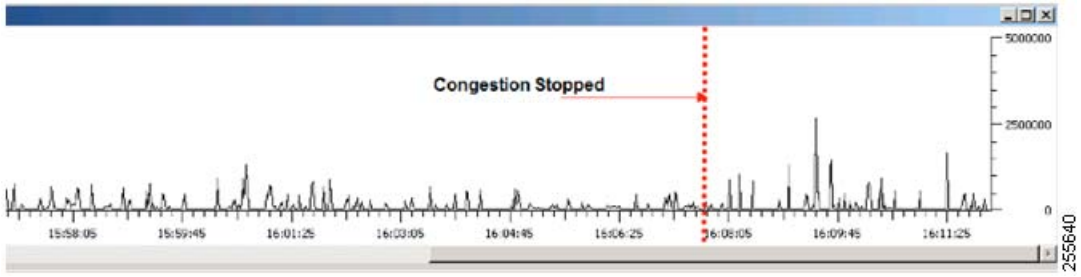
This graph shows the bandwidth utilization measured for one RDP session on the UCS server. Similar to PCoIP, the RDP session shown below is using close to the peak bandwidth measured in bandwidth characterization section earlier and adapts down as congestion increases.

Figure 67 Bandwidth Utilization before and after congestion start with RDP



It is also worth noting that as the congestion traffic is reduced, as with PCoIP, RDP session also quickly ramps up its bandwidth usage and is also above 2.5Mbps within a few seconds.

Figure 68 Bandwidth Utilization before and after congestion stops with RDP



Impact of WAAS Optimization on Cisco Virtual Workspace WAN deployments with View RDP

In the single user RDP session tests above, the minimum bandwidth for a RDP session was determined to be 1.28Mbps using the Cisco KW+ workload. In this test, the objective is to deploy WAAS on either side of the T1 WAN link and characterize the density and performance improvements that WAAS can provide in a Cisco Virtual Workspace deployment using the same workload.

Test Environment and Setup

- View 4.5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory
  - Display protocol: RDP
  - Display Session Characteristics:
    - Screen Resolution: 1366x768 (Large Window)
    - Color Depth: 16bit
  - Windows optimized for Best Performance (All Options checked off)
- Workload Profile: Cisco Knowledge Worker+ profile with optimized antivirus solution from a leading vendor
- Server Profile: UCS B200 M2 with 96G of RAM – server was running at minimal loads during this test
- T1 WAN link was used for this test. Delay of 80ms was injected but no jitter
- WAAS deployed on either end of the WAN link is: WAE-674-K9 running 4.3.1

Application Response Times

The response times for this test with 15 users across a WAN link is well within the success criteria defined in the first column.

Table 18 Application Response Times for WAAS with View RDP

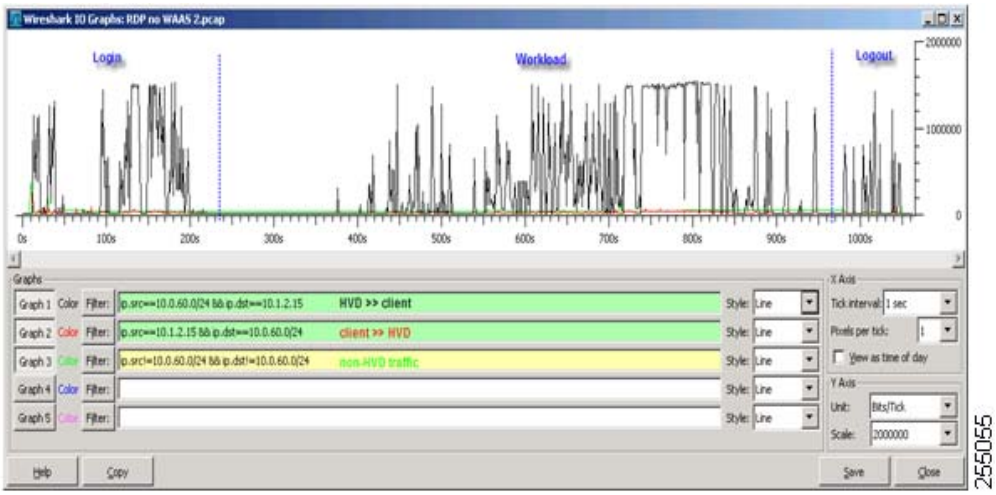
Applications	Success Criteria for Maximum Acceptable Startup Times	Average Startup Times Measured (sec)
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	.96
Outlook	10s	1.9s
Excel	5s	.6s
PowerPoint	5s	.4s
Acrobat	5s	.37s
Internet Explorer	5s	3.6s
Word	10s	7.8s

Summary of Test Results

It was determined earlier that RDP requires a minimum of 1.28Mbps and as a result could only support one user on T1 WAN link. Adding WAAS to this setup for the same workload increased the number of users that can be supported on a T1 link with good UE to 15 users. This was possible due to the 90%+ optimization achieved with WAAS. Though 90% maybe difficult to achieve in real deployments, even a 60% optimization means that a significantly higher number of Cisco Virtual Workspace users can be supported at a branch site with WAAS than without it.

The two graphs below show that going any higher than the # of users possible with the minimum BW value determined earlier results in poor user experience as shown in the application response time table below.

Figure 69 BW utilization with 2 RDP users and without WAAS



255055

Figure 70 Application Response Time for 2 RDP users without WAAS

Timer Name	Average	Min	Max	Standard Deviation
Timer CU: Start	2068	1743	2393	459.619
Timer OL: Start	8806	7680	9932	1592.404
Timer IE: Start	15835	15471	16199	514.774
Timer WD: Start	40423.5	39188	41659	1747.261
Timer EX: Start	2311	2209	2413	144.25
Timer PP: Start	1203.5	1070	1337	188.798
Timer AC: Start	1686	1561	1811	176.777

255056

The figure below shows the response times averaged across the 15 users with WAAS deployed. Note that they are all well within the response time success criteria of 5s/10s defined in an earlier table.

Figure 71 Application Response Time for 15 RDP users with WAAS

Timer Name	Average	Min	Max	Standard Deviation
Timer CU: Start	964.667	760	2364	406.3
Timer OL: Start	1982.267	1730	2239	158.383
Timer IE: Start	3659.067	3165	9962	1743.819
Timer WD: Start	7806.933	6496	24220	4541.02
Timer EX: Start	602.267	505	1197	166.868
Timer PP: Start	409.267	321	982	160.666
Timer AC: Start	373.533	299	1072	195.302

255057

In this next figure we see the data from the WAAS showing the optimization achieved for each of the 15 DV sessions from the branch.

Figure 72 Optimization for 15 RDP users with WAAS

Source IP:Port	Dest IP:Port	Peer Id	Applied Policy / Bypass Reason	Connection Start Time	Open Duration (hh:mm:ss)	Org Bytes	Opt Bytes	% Comp	Classifier Name
10.1.2.15:19344	10.0.60.109:3389	vri-wae-1-waas		19-Apr-11 17:32	0:13:10	23.8819 MB	2.7751 MB	88%	MS-Terminal-Services
10.1.2.15:19351	10.0.60.97:3389	vri-wae-1-waas		19-Apr-11 17:32	0:12:57	24.2096 MB	1.7913 MB	93%	MS-Terminal-Services
10.1.2.15:19357	10.0.60.100:3389	vri-wae-1-waas		19-Apr-11 17:32	0:12:45	25.0142 MB	1.7318 MB	93%	MS-Terminal-Services
10.1.2.15:19363	10.0.60.99:3389	vri-wae-1-waas		19-Apr-11 17:32	0:12:32	24.2054 MB	1.8009 MB	93%	MS-Terminal-Services
10.1.2.15:19372	10.0.60.102:3389	vri-wae-1-waas		19-Apr-11 17:32	0:12:20	25.0199 MB	1.7989 MB	93%	MS-Terminal-Services
10.1.2.15:19378	10.0.60.106:3389	vri-wae-1-waas		19-Apr-11 17:33	0:12:5	21.7068 MB	1.5317 MB	93%	MS-Terminal-Services
10.1.2.15:19384	10.0.60.107:3389	vri-wae-1-waas		19-Apr-11 17:33	0:11:52	24.4126 MB	1.7208 MB	93%	MS-Terminal-Services
10.1.2.15:19390	10.0.60.104:3389	vri-wae-1-waas		19-Apr-11 17:33	0:11:39	24.739 MB	1.7283 MB	93%	MS-Terminal-Services
10.1.2.15:19396	10.0.60.108:3389	vri-wae-1-waas		19-Apr-11 17:33	0:11:26	23.6802 MB	1.8149 MB	92%	MS-Terminal-Services
10.1.2.15:19402	10.0.60.103:3389	vri-wae-1-waas		19-Apr-11 17:34	0:11:13	24.3431 MB	1.8305 MB	92%	MS-Terminal-Services
10.1.2.15:19408	10.0.60.95:3389	vri-wae-1-waas		19-Apr-11 17:34	0:10:59	13.764 MB	1.4088 MB	90%	MS-Terminal-Services
10.1.2.15:19414	10.0.60.105:3389	vri-wae-1-waas		19-Apr-11 17:34	0:10:38	24.289 MB	1.8457 MB	92%	MS-Terminal-Services
10.1.2.15:19420	10.0.60.98:3389	vri-wae-1-waas		19-Apr-11 17:34	0:10:25	25.2784 MB	1.7708 MB	93%	MS-Terminal-Services
10.1.2.15:19431	10.0.60.96:3389	vri-wae-1-waas		19-Apr-11 17:35	0:10:10	24.7778 MB	1.7834 MB	93%	MS-Terminal-Services
10.1.2.15:19437	10.0.60.101:3389	vri-wae-1-waas		19-Apr-11 17:35	0:9:59	25.7576 MB	1.8174 MB	93%	MS-Terminal-Services

255058

The figure below shows the aggregate amount of RDP session traffic and optimization achieved.

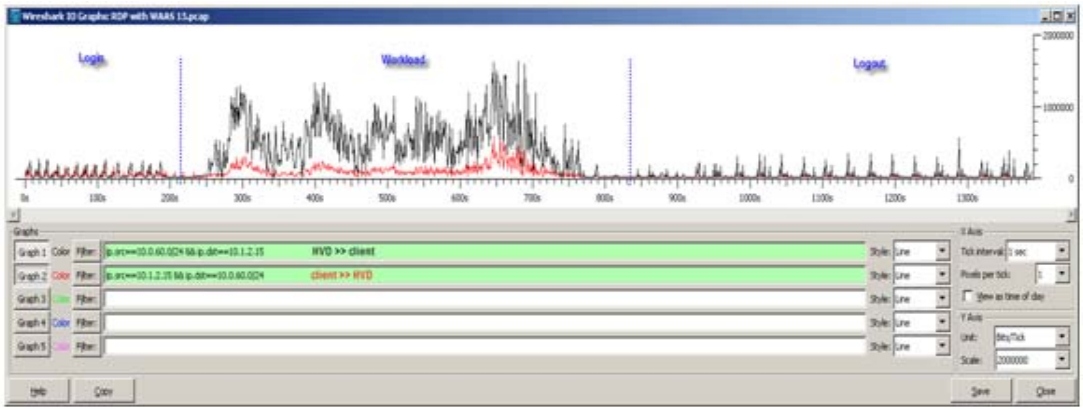
Figure 73 Aggregate optimization across all 15 users

Application	Original Traffic (Excludes Pass-Through)	Optimized Traffic (Excludes Pass-Through)	Pass-Through Traffic	Reduction (%)	Effective Capacity
All Traffic	416.793 MB	58.791 MB	105.852 KB	85.89	7.08 X
Backup	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Content-Management	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Directory-Services	44.900 KB	46.354 KB	0 Bytes	0.0	1.0 X
Email-and-Messaging	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Enterprise-Applications	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
File-System	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
File-Transfer	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
MAPI Reserved Connections	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Other Traffic	4.304 MB	2.289 MB	96.410 KB	46.81	1.88 X
P2P	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Printing	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Remote-Desktop	401.400 MB	47.307 MB	0 Bytes	88.21	8.48 X
Replication	46.174 KB	49.687 KB	0 Bytes	0.0	1.0 X
SQL	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
SSL	495.282 KB	492.822 KB	3.063 KB	0.49	1.0 X
Storage	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Streaming	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Systems-Management	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
Version-Management	0 Bytes	0 Bytes	0 Bytes	0.0	1.0 X
WAFS	530.963 KB	296.892 KB	900 Bytes	44.08	1.78 X
Web	9.998 MB	8.330 MB	5.5 KB	16.68	1.2 X

255059

The next figure shows the bandwidth utilization across the WAN link for non-optimized and other session traffic during login, workload and logout phases of the session.

Figure 74 Non-optimized traffic on the WAN link



Key Takeaways

- In summary, the key takeaways from the network characterization results outlined above are as follows:
- Minimum bandwidth required for PCoIP and RDP with the specified workload is 320kbps and 1.28Mbps respectively. The peak bandwidth consumed by the same workload is 3.6Mbps for PCoIP and its greater than 2Mbps for RDP. This data can be used in sizing WAN links and for enabling QoS polices on these links.
- Certain functions or features within an application may cause peak bandwidth consumption though the application as a whole may not consume as much. For example, slide show mode in PowerPoint has the highest BW impact in the specified workload.
- Cisco Unified Personal Communicator 8.5 in deskphone mode does not have a significant BW impact however PowerPoint and Outlook are the biggest bandwidth consumers in the specified workload.
- WAAS optimization for RDP increased the number of users with good UE from 1 to 15 with 90% optimization. If customers can achieve even 60% optimization with WAAS, it would still be significantly higher than without WAAS.

Rich Media Application Characterization

This section focuses on characterizing various Cisco Rich Media applications so that these applications can be made available to users in a virtual desktop deployment. The following three aspects will be covered here:

- High level summary of deployment profiles tested
- Validation methodology
- Detailed test results

Summary of Results

In this section, a high level summary of the applications characterized across the end-to-end Cisco Virtual Workspace system are provided in the Table 19 below.

Table 19 Summary of Applications

Objective	Server Model	Storage	Desktop Virtualization Profile	HVD Profile
Scale and Performance characterization of Cisco Jabber for Windows with VMware View	Cisco UCS B200 M3 with 384 GB of memory	VSPEX (EMC VNX Series)	VMware View 5.1 on ESXi 5.1	Microsoft Windows 7 32-bit with 2 GB of memory
Scale and Performance characterization of Cisco Contact Center - CTIOS Agent	Cisco UCS B230 M2 with 256 GB of memory	NFS on NetApp FAS 3170	N/A - See test profile for more detail.	Microsoft Windows 7 32-bit with 2 GB of memory

Validation Methodology

The methodology used for doing application characterization is same as that of single server characterization and so please refer to that section for more details.

Detailed Test Results

A detailed analysis of the test results and the associated profile and objectives are provided in this section.

Scale and Performance Characterization of Cisco Jabber for Windows with VMware View on Cisco UCS B200 M3

With Cisco Virtual Workspace (VXI) Smart Solution, Cisco Jabber for Windows is now integrated into Cisco's end-to-end desktop virtualization solution that spans Cisco data center, network and collaboration solutions and based on VMware View.

Cisco Jabber enables an enterprise working model that allows users to collaborate from anywhere, any time using different types of devices such as laptops, desktops, tablets and other mobile devices. Cisco Jabber provides enterprise users with an enhanced collaboration experience by integrating presence, instant messaging (IM), desktop sharing, audio telephony, video telephony and web conferencing into a single software client that runs on the user's physical or virtual desktop, laptop or mobile device. For virtual environments, Cisco Jabber for Windows is available for hosted virtual desktops (HVD) deployed using VMware View. Enterprise users now have the flexibility of using Cisco Jabber from within their virtual desktop session or use locally installed Cisco Jabber on their tablets or smartphones when mobile.

For telephony in virtual environments, Cisco Jabber offers two deployment options, both of which prevent media from hair pinning through the data center. The first option is to use Cisco Jabber running within a virtual desktop to control a physical phone, similar to how one uses Cisco Jabber in a physical desktop to control an external phone. Second option is to use Cisco Jabber to control Virtual Experience Media Engine (VXME) running on user endpoints they use to access virtual desktops. An end-user places calls using Cisco Jabber running on their virtual desktop session and point-to-point media is established between the user's endpoint and other telephony endpoints without the need for a physical phone.

For more details on Cisco Jabber integration into Cisco Virtual Workspace (VXI) Smart Solution, please refer to the Cisco Validated Design for the solution located here: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/VXI/CVD/VXI_CVD_VMware.html

A fundamental consideration when deploying any new application in virtual desktop environment is the impact of that application on the overall desktop load. The cumulative impact of all applications on the desktop and how they are used by each user has a bearing on the shared compute, storage and networking resources in the data center. Therefore when a new application is made available to the users on their desktops, the shared resources that may have been sized based on a different application set must be revaluated to understand the impact of this new application on the shared data center resources. In a large deployment, the impact could be significant depending on how users use the application. For example, if a majority of users start work at a certain time and they all have the pattern of launching their presence and IM application first, then it is important to have a good understanding of the compute, network and storage I/O impact this user behavior has on the shared resources. Adjustments to the shared resources maybe required in order to ensure a success deployment with the application in question. At a minimum, it is important to understand the impact so as to confirm that the current shared virtual resources are sufficient to accommodate the needs of the new application. Otherwise, the users could incorrectly attribute any user experience issues they see as an issue with the application itself. Therefore a new deployment of Cisco Jabber, including migrations from similar applications, should involve an assessment of the application's impact to shared resources.

A first step in this assessment is to understand the incremental impact of adding Cisco Jabber as an application on shared data center resources. First of these shared resources is the compute on the server hosting the desktops or desktop sessions with Cisco Jabber. An enterprise will typically size their servers to accommodate a given number of users so ideally, the assessment with the single application to understand the resource impact at the server level, should also be done with the same density of users. Based on the data from the single server tests, Cisco Jabber resource needs per user can be calculated. The per-server and/or per-user resource utilization data can now be extrapolated to size a Cisco Jabber deployment of any size. The per user data provides the IT administrator with the flexibility to adjust the sizing and extrapolation based on factors in their environment – for example, the IT administrator can assume that only 20% of the users will be using Cisco Jabber simultaneously and if so, the above per-user Cisco Jabber resource data can be used to estimate/adjust the sizing based on 20% of the users using Cisco Jabber simultaneously rather than all users.

When characterizing a single application, the resource impact depends on how the users use the application and the features and capabilities they use. For example, if users at the end of the day typically disconnect from their desktop and leave Cisco Jabber running, the resource impact of many users logging into their desktop, the next day morning, should be less than if they had to start Cisco Jabber first. It is also important to identify specific features in the application that may be particularly resource intensive. One example could be logging or similar features enabled for troubleshooting or monitoring purposes. Logging could increase the I/O load from the desktop and therefore have a greater impact on the storage subsystem. It could also impact the CPU and memory resources that can lower the number of users supported on a given server. Therefore the addition of new applications to a desktop should be done with a good understanding of how the users use the application and the application features being used – together they define the usage profile or workload on the virtual desktop from a single application perspective and could have a bearing on the overall scalability of the deployment from a data center compute, network and storage perspective. Accurately sizing these resources is key to minimizing user experience issues that can impact the overall success of the deployment. Therefore, for any application including Cisco Jabber, any data used for estimating resources needs should be collected with a Cisco Jabber usage profile that reflects, as closely possible the user base that will use Cisco Jabber in production.

Potential changes to the standard desktop configuration are also an important consideration when introducing a new desktop application as it may have CPU, memory and disk requirements than what is currently used. This is particularly important in a virtualized environment with shared compute and storage resources, unlike physical desktops or laptops with dedicated resources. For Cisco Jabber, the minimum requirements when running it in a virtual desktop are: 1vCPU, 2GB of memory and 256MB of disk. See Cisco Jabber data sheet for additional details:

http://www.cisco.com/en/US/prod/collateral/voicesw/ps6789/ps6836/ps12511/data_sheet_c78-704195.html

Another consideration is the application usage pattern across multiple users and the potential peaks in resource usage that this may result in – for example, impact of many users launching and logging into the application at the start of a work day. It is important to ensure that the shared resources can handle periods of peak application usage so that there is minimal impact to user experience.

With the above considerations in mind, testing was performed in the Cisco Virtual Workspace (VXI) Smart Solution to characterize the resource impact of Cisco Jabber for Windows from a compute, storage and network perspective . The testing was done across the end-to-end Cisco Virtual Workspace (VXI) Smart Solution with Cisco Jabber running on 150 virtual desktops deployed on a Cisco UCS B200 M3 using VMware View. The usage profile used in the testing is defined in the Workload Profile section of Table 3. The profile was limited to IM and presence with 150 Cisco Jabber users logging in to Cisco Jabber, loading 200 contacts, and sending and receiving presence updates and instant messages at given rate per user. Testing and results with Cisco Jabber placing and receiving calls should be available in a future release of Cisco Virtual Workspace (VXI) Smart Solution. Note that Cisco Jabber for Windows can be deployed as an on-premise solution or as a cloud based service with the Cisco backend infrastructure hosted in the cloud but the on-premise solution was used in this testing with the Cisco Jabber infrastructure deployed in the same enterprise data center as the Cisco Jabber users.

Though characterizing the application by itself is an important first step when planning for a large virtual desktop deployment, users use multiple applications on their desktop and the overall impact of the application with a more comprehensive desktop workload is still necessary to reflect what happens in production. The overall resource needs of the new application is expected to be less with a comprehensive workload because the simultaneous use of the same application by all users on a server is expected to be less and therefore, less resource utilization by any single application. Results from testing done with a comprehensive (Cisco Knowledge Worker+) desktop workload with Cisco Jabber and other application are also included in the Single Server Scalability Section of this document. However, the per-application data provided here is key to having a detailed understanding of the application and its potential impact to shared resources and therefore the impact of the application to the overall deployment.

In the next section, the results from the Cisco Jabber Application characterization testing done in a VMware View environment are provided.

Validation Overview and Results – VMware View

The goal of this testing is to characterize the scale and performance of Cisco Jabber application deployed on 150 Windows desktops hosted in the data center. For the testing, 150 VMware View virtual desktops were deployed on a Cisco UCS B200 M3 server with 384GB of memory. Cisco Jabber for Windows client was installed on desktops running Windows 7 32b, each with 2GB of memory and 1vCPU.

Test was started by using a Test Tool representing the end users to initiate and login into 150 VMware View desktops. As each user logs into their VMware View desktop, each user launches and logs into Cisco Jabber client installed on the desktops. The test tool then executes the remaining portion of the Cisco Jabber-only workload (see Test Configuration and Setup section below) for a minimum of 2 hours and represents 150 users in steady state use of Cisco Jabber. Once the workload has been running for a while, the process of logging off the users from their desktop is initiated. During the desktop logout

stage, users also log off and quit the Cisco Jabber client running on the desktop. The resource utilization data is collected through all stages of Cisco Jabber use, including desktop session launch and login by running resxtop on the server that collects the utilization data directly from the hypervisor using a polling interval of 5s.

The performance graphs for the data collected from the Cisco UCS server are provided in the Performance Charts section below. The charts shows the Cisco Jabber resource utilization for 150 desktops from a compute, network and storage perspective through different stages of Cisco Jabber use - Launch and Login, Steady State Use and Desktop session logout. The data from the performance charts are also summarized in the table below. The setup and workload/usage profile used in the testing are also outlined in the Test Configuration and Setup section below.

Table 20 Resource Utilization on a Cisco UCS B200 M3 server with 150 VMware View desktops running Cisco Jabber

	Launch & Login	Steady State	Desktop Session Logout
CPU Utilization-Avg.	28.58	20.84	20.28
CPU Utilization-Peak	41.15	29.83	48.25
Memory Allocated (%)	-	82.87	-
Read-Avg	78.74	28.28	27.41
Read-Peak	305.60	873.78	112.44
Write-Avg	351.84	330.20	330.50
Write-Peak	731.00	832.50	801.39
Read-Latency-Avg.	0.69	0.55	0.71
Read-Latency-Peak	1.30	1.99	1.38
Write-Latency-Avg.	0.80	0.80	0.74
Write-Latency-Peak	1.05	3.12	1.30

The data shows that Cisco Jabber uses approximately 20% of the server's compute resources during steady state workload stage when all users are using their desktop per the workload profile defined in the Test Profile section below. During the launch and login stage, CPU utilization on the server is at ~30% (average) and 40% (peak). This is for ~10 minutes when the 150 users are launching and logging into their Cisco Jabber client at the start of the workload.

From a memory utilization perspective, approximately 80% of the available memory on the server was allocated to the 150 desktops with 2GB of memory per virtual desktop. The UCS server used in the test was deployed with 384GB of memory. The utilization of 80% represents the memory allocated to 150 virtual desktops, along with memory used by the ESXi hypervisor and virtualization overheard. The actual memory usage will depend on the workload and should be monitored in production at the UCS server level to ensure that there is memory available for supporting the desktop users running on that server. For environments that use memory over-subscription, the overall memory deployed on the server could be lower based on observed usage.

From a storage perspective, the average I/O load generated by 150 Cisco Jabber users for the given workload profile is approximately 30 peak IOPS and 325 write IOPS for a combined total of 355 average IOPS. Peak I/O load generated is approximately 300 peak read IOPS and 825 peak write IOPS, for a total of ~1125 peak IOPS. Excluding the peak read I/O data from steady state as it is momentary (see

Performance charts below) and considering that virtual desktop workloads are typically write I/O intensive during steady state (read/write ratios as high as 10/90) so assuming this to be a temporary glitch in the test environment.

The I/O activity in the logout stage involves logging off from Cisco Jabber server, closing Cisco Jabber application and logging off from the virtual desktop.

The I/O load generated by a Cisco Jabber workload is consistent with the I/O profile of a virtual desktop workload in terms of being peak read I/O intensive during login and write I/O intensive (relative to read I/O) during all stages of use. Based on the server level I/O data for 150 users, the per user Cisco Jabber I/O requirements can be estimated as 1/2 for average read/write IOPS and 1/8 for peak read/write IOPS.

I/O latency experienced by Microsoft Windows OS running on the desktops is well below the acceptable threshold of 20ms (average) throughout the test.

Based on the above data, resource usage per desktop using Cisco Jabber can be calculated and used for planning a deployment of any size. Note that to ensure the accuracy of any estimation used in planning, it is best to validate the estimations through proof-of-concept type testing in the enterprise environment where it will be deployed.

Table 21 Compute, Storage and Performance Requirements for a single desktop running Cisco Jabber

Compute	Average = ~62 MHz	Derived using the following calculation: • Cisco UCS B200 M3 = 2 x 8 core x 2.9 GHz = 46.4 GHz of compute capacity • Average CPU utilization measured (table above) = 20% = .20x 46.4GHz = 9.3GHz • Average CPU cycles needed per desktop = 9.3GHz/150 = 62 MHz
Memory	2GB per user	Assuming no memory over-subscription
Storage I/O	Average = ~1/2 for Read/Write IOPS Peak = ~1/8 for Read/Write IOPS	Derived using the following calculation: • Average = ~30R/325W IOPS/150 users = ~1R/2W IOPS/user • Peak =~300R/825W IOPS/150 users = ~1R/8W IOPS/user

The remainder of the section provides a detailed overview of the test setup, workload and results. The results include performance charts and Cisco Jabber response times for the testing done with 150 virtual desktops deployed on a Cisco UCS B200 M3 server.

Test Configuration and Setup

This table below provides configuration, environment and setup details used in the testing.

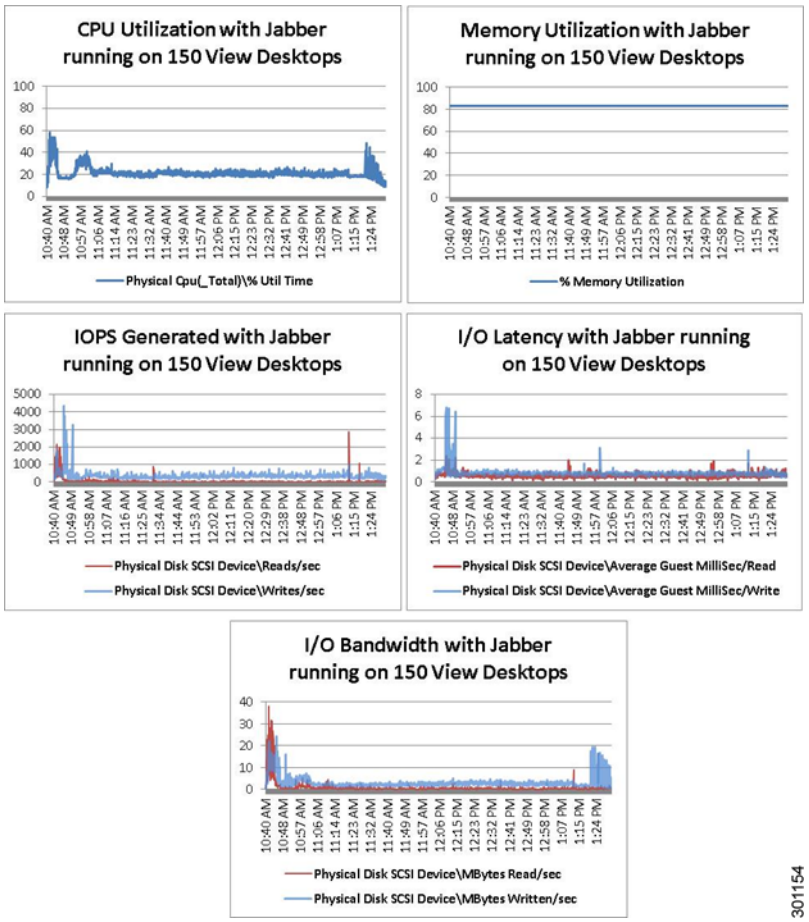
Table 22 Configuration and Setup used in Cisco Jabber testing across 150 VMware View desktops

Desktop Virtualization	VMware View 5.1
UCS Server	UCS B200 M3 with Dual Eight Core Intel®Xeon® CPU E5-2690@ 2.9GHz with 384 GB of memory
Hypervisor	VMware ESXi 5.1
Storage	VSPEX (EMC VNX Family)
Virtual Desktop Configuration	Windows 7 32-bit desktops with 2GB of RAM and 20G disk, 1 vCPU, No memory reservation
Cisco Jabber for Windows	9.1.3

Desktop Virtualization	VMware View 5.1
Workload Profile	Test was conducted with Cisco Jabber being the only application being used on the desktop. For this reason, the workload profile is same as the Cisco Jabber usage profile outlined as follows: • Total Contacts Per User ((The contacts are mutual friends of each other) = 200 • Online Contacts during testing = 150 • Offline Contacts during testing = 50 • Cisco Jabber workload on each desktop can be summarized as follows: – Login to Cisco Jabber – Desktop experiences State Changes at a rate of 8 per hour per user (either sent by the user or received from other users) – Initiate Instant Message chat sessions to 4 other users – Send Instant Messages on each of the above 4 chat sessions at a rate of 5 per hour per user – Message Sent: "OMG! The quick brown fox jumped over the lazy brown dog!" • The above workloads runs on all 150 user desktops • The exact steps performed during testing are outlined below: – Launch and Login to 150 VMware View virtual desktops – Wait until Desktop Login phase completes – Start the workload using test tool; tool will stagger the start of the workload so that the workload is randomized across the 150 desktops – Execute Cisco Jabber workload described above – Allow the test to run for a minimum of 2 hours in Workload Steady State – Logout of Cisco Jabber, Logout of virtual desktop that closes out Cisco Jabber application
Data Collection & Test Tools	• Workload Generation - Scapa Test Performance Platform (TPP) • Resxtop with a polling interval of 5s is used to measure the hypervisor resource usage metrics • End user response times measured using Scapa • Data is captured and graphed for Cisco Jabber Launch & Login, Steady State use and Desktop session logout (with logout and closing of Cisco Jabber) stages of Cisco Jabber use

Performance Charts

Figure 75 Performance Charts for a Cisco UCS B200 M3 with 150 VMware View desktops running Cisco Jabber



Application Response Times

The table below shows that the response time experienced by 150 users were well within the established success criteria of 5sec.

Table 23 Response Times for 150 users on Cisco UCS B200M3 with View5.1/ESX5.1/PCoIP/EMC VNX

Application Running on the HVD	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 150 VMware View users on UCS B200M3
Cisco Jabber 9.1.3 for Windows	5s	1.6s

Summary

Though the events in the workload are randomized, the data collected from this testing is with all 150 users actively using a single application – Cisco Jabber. In production virtual desktop deployments, users are using different applications and at different times so percentage of users actively using Cisco Jabber

at any given time could be less than what we have assumed for this test. Therefore the data provided here shows the upper limits of resource utilization for 150 users using Cisco Jabber as defined in the Cisco Jabber workload profile. For this reason, enterprises should attempt to evaluate their usage model and adjust the sizing accordingly - this data provides a starting point for the sizing exercise for a given workload with 150 users actively using Cisco Jabber on their VMware View desktops.

Scale and Performance Characterization of Cisco CTI OS on Cisco UCS B230 M2

A fundamental aspect of deploying Cisco Virtual Workspace in a call center environment is the virtualization of agent desktops. In order to virtualize and host the agent's desktop from the data center, the compute, storage and networking needs of the agent must be well understood. The resource needs will depend on how the agents use their desktop in terms of their usage profile and the type of applications used. Call center users will be fundamentally different from other desktop users in the same Enterprise due to the unique nature of their jobs. Call center desktop users are often characterized as Task Workers to indicate a lighter workload while the average Enterprise user is referred to as Knowledge worker to imply a heavier workload. Knowledge workers may use several applications at a time, from Microsoft Office applications to collaborating with their peers using Cisco Jabber or Cisco WebEx, to browsing the web, downloading documents etc. Call center workers may also use the same desktop applications but when they do, they might only use one or two applications at any given time and may not multi-task to the extent that a Knowledge Worker does. But more importantly, the primary application they use could be a customized application in order to do their job. The differences in the workload defined by the application set and the usage profile is an important distinction that has bearing on the shared virtualization resources required to support a call center agent desktop deployment. For any deployment, any data used for planning purposes should be based on a workload that best represents the workload of the users in production. Otherwise sizing estimations for compute, storage and network may completely miss the mark for the deployment in question.

To aid in capacity planning for a contact center deployment, in this section, we focus on Cisco contact center environment and specifically on Cisco agent desktop software that an agent will primarily use for accepting calls and working with customers. Therefore it is important to understand the compute, storage and network requirements of the one application that the agent will use throughout their shift. A comprehensive workload with other applications, such as the Cisco KW+ workload used in other scale and performance testing, was not used in this testing for two reasons. First, there is a high degree of variability in the application set used by agents in call center environments. Secondly, the applications used are entrenched applications that are heavily customized and require extensive backend infrastructure that cannot easily be replicated in a test environment. Therefore, the testing covered in this section strictly focusses on Cisco agent desktop software, namely CTI OS and provides resource utilization that can be used as a starting point for assessing the overall resource needs of a virtualized agent desktop deployment.

Another important consideration in call center environments is the collective impact of how the call center operates such as whether they follow shift based work or follow the sun type working working models. These transition points are important for capacity planning, as they are also periods of peak resource usage when desktops are powered on in preparation for the new shift. Another period of peak activity is at the start of a shift when all are launching applications and logging into their contact center environment to start taking calls. Just as these transition events can impact the back end call center server infrastructure, they can also impact the shared resources in a virtualized environment. In call center deployments, it is particularly important to plan for these login or boot storms since they can occur more frequently with every shift change. For this reason, the usage profile used in this testing was defined such that it included a period of peak usage to reflect shift change type events in a call center environment.

Results Summary

As stated earlier, the objective of this testing in the Cisco Virtual Workspace system is to provide resource utilization data for virtual desktops running Cisco CTI OS agent software which can be used in capacity planning a virtual agent desktop deployment based on Cisco contact center solution. For the testing 120 virtual desktops running Cisco CTI OS were deployed on a Cisco UCS B230 M2 with 256GB of memory. Each desktop was deployed as a Windows 7 machine with 2GB of memory each. Due to the memory allocation per desktop, approximately 90% of the server's available memory was allocated to the 120 desktops. Note that the 120 desktops deployed on the server for this testing does not reflect the maximum number of users this server can support. Determining the maximum scalability of the UCS server was not the objective of the test. Instead, the objective was to characterize a virtualized Cisco CTI OS application to determine the performance impact on shared virtualization resources. For this purpose, a server with significant load was needed. Loads of 120 users were used based on the 'allocated' memory being 90% based on a 2GB per desktop configuration.

From a CPU perspective, Cisco CTI OS has minimal impact on server's CPU resources during steady state workload stage when all agents are using their desktop per the workload profile defined below. CPU utilization is less than 20% during steady when all 120 agents are actively using Cisco CTI OS to receive calls and talking to customers. However, CPU usage does peak to 99% utilization for a brief period of time, approximately 30s, when all users are launching and logging into Cisco CTI OS. This is to be expected and represents an application level storm, with all users attempting to come up almost simultaneously.

From a storage perspective, the I/O requirements during peak and steady state workload stages are approximately 1900 and 500 IOPS respectively with this workload. Read IOPS peaks to 800+ IOPS during peak usage when agents are launching and logging into CTI OS and stays well below 100 IOPS for the remainder of the time. Write IOPS also peak during peak usage to 1100+ IOPS but stays steady at approximately 400 IOPS until logout where it again peaks to around 1100 IOPS. Logout stage involves logging off from the CTI OS server and closing the CTI OS application running on the desktop. Also, I/O latency experienced by the Guest OS (Microsoft Windows) on the desktops is well below the acceptable threshold of 20ms throughout the test.

From a network perspective, peak bandwidth (BW) usage is 30 MB/s (240Mbps) for storage traffic and 250 Mbps for other types of network traffic. Peak bandwidth usage coincides with the peaks in CPU and I/O and occurs during the launching and logging in of CTI OS on 120 desktops. However during steady state workload, CTI OS on 120 desktops requires only 2 MB/s (16Mbps) of storage and 20 Mbps of other network traffic. Logout also shows an increase in utilization of approximately 12.5MB/s (100Mbps) for storage and 100Mbps for other network traffic. Note that the network bandwidth utilization does include the BW associated with the audio calls as these calls will never be seen by the agent desktop and therefore not in the server level bandwidth measurements. Also, the tests were done directly from within the desktop and therefore is also no desktop virtualization display traffic that is typically transported across the network to a user device used to access the virtual agent desktop. To size the bandwidth requirements for the display traffic associated with exporting the agent desktop running Cisco CTI OS client, it is best to do this by measuring the bandwidth a single session as the agent uses their desktops, specifically for the launching applications, logging in and taking calls. Note that display protocols are adaptive and proprietary and can change with network conditions. Therefore it is best to assess the bandwidth requirements with the network conditions that the agents will typically experience. For example, if the agents are located in a branch site with the desktops in a central data center and the latency on the WAN link is 80ms, the bandwidth per session with good experience for the branch site may not be the same as a campus user connected via a LAN. Please refer to Network Characterization section for more details on bandwidth sizing in a virtual desktop deployment.

Lastly, it is important to stress that any variations in the Cisco CTI OS usage profile or workload used in the testing can change the resource utilization. For example, the Busy Hour Call Attempts (BHCA) for an agent desktop and the number of skills group that are enabled for the agent are key factors that can increase the resource needs of a Cisco CTI OS based virtual desktop deployment.

The above discussion on the overall resource utilization of 120 agent desktops running CTI OS are summarized in the following table. The usage profile for Cisco CTI OS used in this testing is outlined in detail in the next section.

Table 24 Resource Utilization on a Cisco UCS B230 M2 server with 120 virtual desktops running Cisco CTI OS

CPU Utilization	Peak = 99% Average = 20%	Peak occurs when all 120 desktops are launching CTI OS and logging in
Memory Utilization	Average = 90%	This reflects the total memory allocated by ESXi hypervisor to 120 agent desktops with 2G of memory each
Storage	Peak I/O = ~2000 (Read/Write=900/1100) Average I/O = 500 (Read/Write = 100/400)	Peak I/O occurs when CPU also peaks as outlined above Average I/O is during steady state workload stage when agents are using their desktop per the workload definition in the next section
Network	Peak Network BW Utilization = ~500 Mbps Average Network BW utilization = ~50 Mbps	Peak BW utilization occurs when CPU and I/O also peaks as outlined above Bandwidth utilization includes all network traffic, including storage

Based on the above data, resource usage per agent using Cisco CTI OS can be derived and used for planning a deployment of any size. Note that to ensure the accuracy of any estimation used in planning, it is best to validate the estimations through proof-of-concept type testing in the Enterprise environment where it will be deployed.

Table 25 Compute, Storage and Performance Requirements of a single Cisco CTI OS agent desktop

Compute Required per CTI OS Agent	Average = ~80 MHz	Derived using data from previous table: • Cisco UCS B230 M2 = 2 x 10 core x 2.4 GHz = 48 GHz of compute capacity[1] • Average CPU cycles available per desktop = 48 GHz/120 = 400 MHz • Average CPU cycles used by CTI OS during steady state use for the usage profile used in this testing = 20% of 400MHz = 80MHz • Data reflects the overall needs of the agent desktop running Microsoft Windows and Cisco CTI OS client
Memory Required per CTI OS Agent	Average = ~550MB	Measured directly at the Guest OS level and reflects the overall needs of the agent desktop running Microsoft Windows and Cisco CTI OS client
Storage I/O Performance Required per CTI OS Agent	Peak = ~15–20 IOPS (Read/Write= ~8/9) Average = ~5 IOPS (Read/Write = ~1/4)	• Data reflects the overall needs of the agent desktop running Microsoft Windows and Cisco CTI OS client • Derived using data from previous table: – Peak = ~2000 IOPS/120 users = ~17 IOPS/user – Average = ~500 IOPS /120 users = 4+ IOPS/user
Network BW Required per CTI OS Agent	Peak Network BW Utilization = ~5 Mbps Average Network BW utilization = ~500 kbps	• Data reflects the overall needs of the agent desktop running Microsoft Windows and Cisco CTI OS client • Derived using data from previous table – Peak = ~500 Mbps/120 users = ~4 Mbps+/user – Average = ~50 Mbps /120 users = ~420 kbps/user

[1] The overall compute performance of a server, particularly in the newer generation processors, is not strictly a factor of clock speed and number of cores. The processor architecture, in terms of memory speeds and throughput, the amount of L1, L2 processor cache, the number and speed of connections between CPU sockets are all factors that can improve the overall compute performance. The calculation used here is nevertheless a straightforward method to quantify the minimal performance that can be expected from a server.

The remainder of the section provides detailed information on the deployment profile, workload and other configuration/setup information. The performance data measured at the server level using resxtop with a polling interval of 5s are also provided below.

Detailed Performance Results

This section provides a detailed overview of the test setup and results in terms of the configuration and performance charts with Cisco CTI OS client running on 120 desktops, deployed on a Cisco UCS B230 M2 server.

Test Profile

Table 26 provides configuration, environment and setup details used in this testing.

Table 26 Configuration and Setup used in Cisco CTI OS testing across 120 virtual desktops

Desktop Virtualization	N/A as test was conducted by running a script directly on the virtual desktop - data is independent of the desktop virtualization solution
UCS Server	UCS B230 M2 with Dual Ten Core Intel® Xeon® CPU E7-2870@ 2.4GHz with 256GB of memory
Hypervisor	VMware ESXi 5.0U1
Storage	NetApp FAS 3170
Virtual Desktop Configuration	Windows 7 32b desktops with 1vCPU, 2G of RAM and 20G disk; No memory and CPU reservations for the agent desktop virtual machines
Cisco Contact Center	CTI OS Server and Client side software version: 9.0.1

Workload Profile	Test was conducted with Cisco CTI OS as the only application running on a virtual desktop. The usage of profile of the application defines the workload on the desktop and this defined as follows for the automated workload used to perform the tests: • Agent launches CTI OS Client • Agent starts Microsoft Internet Explorer • Logs into CTI OS server • Hits the 'READY' Button on the CTI OS Client UI to indicate to Contact Center that it is ready to receive calls • Contact Center System starts sending calls to agent; agent accepts calls; duration of the calls are anywhere from 1min - 5min during which the agent browsed 3 web pages; agent ends the call • Agent receives next call. Previous step repeats and this repeats itself for the duration of the test (~2 hours) • Agent then toggles 'READY' button to stop receiving calls and logs off when the test ends • Same events occur on all virtual desktops running on the Cisco UCS server • Simulated calls were sent to Contact Center system (to be received by agents) at a BHCA of 9000 calls spread across 200 simulated phones. Each agent takes approximately 1 call every 5min, 12 BHCA per agent and 1440 BHCA across 120 users
Data Collection & Test Tools	• Workload script used to emulate the actions of the agent • Workload script automatically runs when the agent logs in • At the server level, resxtop is used to measure resource usage metrics reported by the hypervisor • Data is captured and graphed for Login, Workload and Logout stages of CTI OS client use by the agent

Performance Charts

Figure 76 Performance Charts for Cisco UCS B230 M2 with 120 Contact Center Agent desktops running Cisco CTI OS client