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

May 3, 2013

Contents

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

Introduction

This performance guide serves as an addendum to the Cisco Virtual Workspace (VXI) Smart Solution 2.7 Performance and Capacity Results Guide for Citrix located here http://www.cisco.com/en/US/partner/docs/solutions/Enterprise/Data_Center/VXI/CVD/VXI_CVD_Citrix.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 Cisco Virtual Workspace system. The primary focus is on characterizing the scalability and performance of 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 Cisco Virtual Workspace (VXI) 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 192G of memory per server with the exception of UCS B230 M2 with 256G of memory.

In the Cisco Virtual Workspace system, the following hypervisors have been validated with Citrix XenDesktop:

- VMware vSphere (ESXi)
- Citrix XenServer
- Microsoft Hyper-V

Citrix XenServer

XenServer 5.6 FP1 and 6.0 has been validated with XenDesktop in the Cisco Virtual Workspace system. XenServer now supports IntelliCache which is a hypervisor based storage optimization solution that uses high performance disks (local) on the server as a cache. IntelliCache can offload the IOPS going to the back-end storage array by serving those requests from the cache hosted on the server's local disks, thereby reducing the overall costs associated with hosted virtual desktop deployments. Cisco UCS B-series and C-series servers both support SSD drives that are well suited for this purpose. For more details on this feature, please refer to the following article (http://support.citrix.com/article/CTX129052). The optimization data from the IntelliCache testing will be available soon in the next update to this Performance Guide.

Note IntelliCache is available with XenDesktop using Machine Creation Services (MCS).

VMware vSphere (ESXi)

ESXi is validated for a number of deployment profiles in the Cisco Virtual Workspace system with sizing and performance data available for both ESXi 4.1 and ESXi 5.0. 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 the Cisco Virtual Workspace Solution 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 Virtual Workspace Solution.

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 HVD 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. Targeted testing was also done in the Cisco Virtual Workspace Solution 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.

Microsoft Hyper-V

Cisco Virtual Workspace also validated Hyper-V 2008 R2 SP1 which includes the dynamic memory management feature that can significantly increase the scalability of your virtual desktop deployments when using Hyper-V. This feature enables a more efficient use of the physical memory by allowing the memory to be allocated to the virtual desktops on an as needed basis. Previously, a desktop was pre-allocated a fixed amount of memory (static) regardless of whether the desktop needed it or not. With dynamic memory allocation, each HVD specifies a maximum and a minimum value for the memory when the desktop is created. The hypervisor then uses this range to allocate more or less memory based on what the desktop needs are. To determine the scalability impact of dynamic memory allocation, testing was done in the Cisco Virtual Workspace system to determine the max density with both static and dynamic memory allocation. The physical memory on the UCS servers were to limited to 96G in both cases.

Virtual Desktops are validated in the Cisco Virtual Workspace system using XenDesktop with Machine Creation Services and Provisioning Server (to a lesser extent) and both are fully supported. Data from XenDesktop 5 through XenDesktop 5.6 are both included in this document.

Deploying desktops using MCS is highly beneficial due to the storage capacity savings it provides especially considering that high storage costs are a hindrance to desktop virtualization adoption. The storage savings come from the individual desktops sharing a common base image called a master or parent image. Each desktop uses a desktop image that consists of:

- Larger shared common base image that is read only
- A 16mb unique identity disk that uniquely defines the desktop
- A unique differential disk that stores any changes made to the desktop

The differential disks are also thin provisioned which maximizes storage use by allocating space only as needed. Using pooled desktops that share the larger common image in this manner can 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. The identity disk and the differential disk uses a fraction of the disk space so the total storage savings are significant. A key trait of this architecture is the separation of the main OS disk from differential disk that captures all changes a user makes. Storage tiering and caching technologies discussed below can address the IO requirements of this model and further reduce costs while improving user experience.

MCS can be used to deploy both persistent(dedicated) desktops where changes to desktop is retained and non-persistent(pooled) desktops where the desktop is refreshed back to the original state at reboot. In the Cisco Virtual Workspace system, validation was done with pooled desktops though other types of desktops are also supported. 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 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 CVD for Citrix. Note that that version (KWP 1.6 or KWP 2.5) of the 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, differential disks associated with each desktop 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 EMC (VNX-series) 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 live migration. 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. Performance data using Atlantis will be available soon in the next update to this Performance Guide.

Please refer to the Performance and Capacity chapter of the Cisco Virtualization Experience Infrastructure Smart Solution 2.6 with Citrix XenDesktop 5.6 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
Scale and Performance characterization of Cisco UCS B200 M3 with Citrix XenApp Hosted Shared Desktops using a Cisco KW+ workload without antivirus	Cisco UCS B200 M3 with 384G of memory	NFS on NetApp FAS 3170	Citrix XenApp 6.5 on ESXi 5.1	Microsoft Windows 7 32-bit with 2 GB of memory and 24G disk; Persistent
Scale and Performance characterization of Cisco UCS B200M3 server with Citrix XenDesktop (FlexPod)	Cisco UCS B200 M3 with 384G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop 5.6 FP1 (MCS) on ESXi 5.1	Microsoft Windows 7 32-bit with 2 GB of memory and 24G disk; Persistent
Impact of deploying Citrix PVS write cache on local SSDs	Cisco UCS B200 M3 with 384G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop 5.6 FP1 (PVS 6.1) on ESXi 5.0 U1	Microsoft Windows 7 32b with 1.5G of memory and 24G disk; Non-persistent
Scale and Performance characterization of Cisco UCS B250M2 server (FlexPod)	Cisco UCS B250 M2 with 192G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop5 on ESXi 5.0; ICA	Microsoft Windows 7 32b with 1.5G of memory and 24G disk; Persistent

Objective	Server Model	Storage	Desktop Virtualization Profile	HVD Profile
Scale and Performance characterization of Cisco UCS B230M2 server (FlexPod)	Cisco UCS B230 M2 with 256G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop 5.5 on ESXi 5.0; ICA	Microsoft Windows 7 32b with 1.5G of memory and 24G disk; Persistent
Scale and Performance characterization of Cisco UCS B250M2 server - Impact of Hyper-V Dynamic Memory	Cisco UCS B250 M2 with 192G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop 5.5 on Microsoft HyperV 2008 R2 SP1	Microsoft Windows 7 32b with 1.5G of memory and 24G disk; Persistent
Storage Optimization (IOPS Offload) using Citrix Intellicache	Cisco UCS B-230 M2 with 256G of memory	NFS on NetApp FAS 3170	Citrix XenDesktop 5.5 on XS 6.0*	Microsoft Windows 7 32b with 1.5G of memory and 24G disk; Persistent

Note Please refer to Citrix Product Documentation and Cisco Virtual Workspace (VXI) Smart Solution 2.7 Release Notes for support related information and caveats:
http://support.citrix.com/proddocs/topic/xendesktop-als/cds-installing-xd5fp1.html
http://www.cisco.com/en/US/customer/docs/solutions/Enterprise/Data_Center/VXI/VXI_RN_CPE.pdf

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

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 Performance and Capacity Results Guide for Citrix 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 Knowledge Worker (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 Performance and Capacity Results Guide for Citrix 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 some exceptions
- No memory ballooning (ESXi) or host swapping ok, i will le
- Average IO Latency less than 20 ms

Application Response Times

Table 2 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

Applications	Success Criteria for Maximum Acceptable Startup Times
Word	5s*
* Testing in previous releases of VXI 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. For other hypervisors, iostat and perfmon are used for XenServer and Hyper-V respectively. 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 Citrix XenDesktop 5.6FP1

When deploying a Cisco Virtual Workspace (VXI) Smart Solution based on Citrix XenDesktop 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 FlexPod infrastructure running Citrix XenDesktop 5.6 FP1 and VMware ESXi 5.1. Results indicate that ~130 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 <1sec with one application having a response time in the 1-2s 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 130 desktops on a Cisco UCS B200 M3 server.

Summary of Test Results

Using the above deployment profile, 130 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 <15ms
- Application Response times <2sec

Test Profile

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

Desktop Virtualization

- Citrix XenDesktop 5.6 FP1
- Connection Protocol – ICA
- Pooled Desktops (Reboot on logoff disabled for ease of testing)

Hypervisor

VMware ESXi 5.1

Virtual Desktop Configuration

- Windows 7 32b desktops with 1.5G of RAM, 24G disk and 1vCPU per desktop
- Persistent desktops (Due to reboot after logoff being disabled)

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+ (Ver 4.25)

- 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

- NAS NFS

- NetApp FAS 3170 with PAM 2 module (512GB of cache)

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 130 users were well within 5sec or the established success criteria for all applications.

Table 3 Response Times for 130 desktops on Cisco UCS B200M3 with XD5.6FP1/ESX5.1/ICA/NetApp Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 130 desktops on Cisco UCS B200M3
Cisco Jabber for Windows (Version 9.13)	5s	0.58s
Outlook'10	5s	1.7s
Word'10	5s	0.73s
Excel'10	5s	0.65s
PowerPoint'10	5s	0.68s
Internet Explorer	5s	0.58s
Acrobat	5s	0.63s

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 130 Citrix desktops running a Cisco KW+ workload are shown in the figure below.

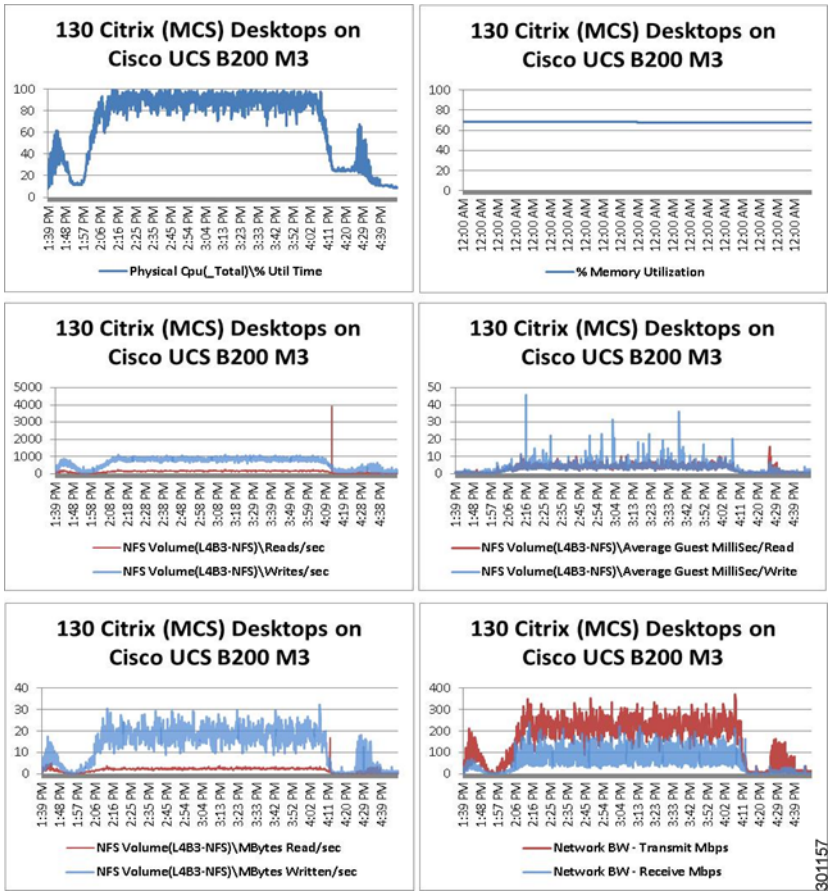
The first chart shows the CPU utilization measured using resxtop with a 5s polling interval during the Login, Workload and Logout stages with 130 desktops on a Cisco UCS B200 M3 server. This chart confirms that we are CPU bound for this profile with a CPU utilization of ~ 90% during steady state use of their desktop by 130 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 384 GB 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 average I/O latency is <5msec – 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.

The last chart shows the network bandwidth utilization which is a combination of the storage traffic and all other traffic sent and received by the desktop and can be a starting point for estimating the bandwidth needs in the data center.

Figure 1 Performance Charts for 175 desktops on Cisco UCS B200M3 with XD5.6FP1/ESX5.0U1/ICA/NetApp Profile



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 130 users

Storage I/O	Steady State	Login	Logout
Read-Avg.	156.02	110.96	27.71
Read-Peak	247.33	210.54	111.25
Write-Avg.	836.95	533.53	241.80
Write-Peak	1105.31	870.59	599.44
Read-Latency-Avg.	4.25	0.84	0.96

Storage I/O	Steady State	Login	Logout
Read-Latency-Peak	10.84	2.04	15.55
Write-Latency-Avg.	4.73	0.80	0.88
Write-Latency-Peak	46.14	2.95	7.14

The I/O data shows the steady state read and write IOPS are approximately 850 write IOPS and 150 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 8 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 Citrix XenDesktop 5.6FP1 and ESXi5.1 and running Cisco KW+ workload is 130 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.

Hosted Shared Desktop (HSD) Scalability on Cisco UCS B200M3 with Citrix XenApp 6.5

With Citrix XenApp Hosted Shared Desktops, Cisco Virtual Workspace (VXI) Smart Solution offers an alternate desktop delivery solution for Enterprise users do need require a dedicated desktop such as Call Center workers that may have a lighter desktop workload. The advancements in Cisco UCS server technology is leveraged by Cisco Virtual Workspace (VXI) Smart Solution to deliver a highly scalable, shared desktop solution based on Citrix XenApp.

Citrix XenApp uses Citrix FlexCast TM technology to deliver both applications and desktops from a centralized data center. For desktop delivery, Citrix XenApp offers Hosted Shared Desktops (HSD), also known as Published desktops or Shared Hosted Desktops. XenApp desktop delivery is very similar to Microsoft Remote Desktop Services or Terminal Services. With HSD, users share a single Microsoft Windows Server and establish independent desktop sessions to the server using ICA. In Cisco Virtual Workspace (VXI) Smart Solution, the Microsoft Windows server is virtual machine hosted in the data center on Cisco UCS servers. Multiple Microsoft Windows servers can be deployed on the same Cisco UCS server to support a large number of users. Typically, HSD offers a highly scalable solution at lower storage, network and server costs than that of an equivalent hosted virtual desktop deployment – however, it is important to note that an Enterprise can deploy both delivery options, as they each address a different use case and a different user base.

When planning for a deployment based on Cisco Virtual Workspace (VXI) Smart Solution and XenApp HSD, it is important to understand the scalability and performance of this delivery model from a compute, network and storage perspective. From a compute perspective, it is important to have an optimal configuration that maximizes the number of users on a server to minimize costs. From a network and storage perspective, it is important to characterize the network and I/O impact of this model so as to size the storage and network needs of the deployment. With these objectives in mind, testing done in Cisco Virtual Workspace (VXI) Smart Solution focusses on determining the scalability of XenApp HSD based deployment on Cisco UCS servers. For this testing, XenApp server VMs running Microsoft Windows 2008 R2 are deployed on a Cisco UCS B200 M3 server with users accessing the shared desktop across the end-to-end system. The scalability of a single Cisco UCS B200 M3 server in terms of the number of users it can support is an important data point that will determine the total number of servers

needed for a given deployment. The storage (I/O, I/O bandwidth, I/O latency) and network related metrics are measured at max scale and can be used for sizing the storage subsystem and the network links.

Based on the testing done in the Cisco Virtual Workspace (VXI) Smart Solution, results indicate that approximately 160 users can be supported on a single Cisco UCS B200M3 using Cisco Knowledge Worker+ (KW+) workload (but no antivirus running on the XenApp server VM). Results also indicate that we are CPU bound for this profile. Note that the Cisco KW+ workload used in this testing maybe more intense than the typical Task Worker type HSD workloads. The optimal configuration to achieve the above scalability is 8 XenApp server VMs, each with 4vCPUs and 16GB of memory. This implies that a single Microsoft Windows server VM can support approximately 20 users each using this workload. Cisco UCS server should be minimally deployed with 144GB ((8 XenAppVMx16GB) + ESXi + virtualization overhead + buffer) of memory. This memory configuration is balanced and performance optimized. Cisco recommends using Cisco UCS VM-FEX technology in XenApp deployments for improved response times though it was not used in this testing.

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 160 HSD users on a Cisco UCS B200 M3 server.

Summary of Test Results

Using the above deployment profile, 160 users 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= ~35% (Installed memory was far higher than needed)
- Average I/O Latency <20ms
- Application Response times <4sec

Test Profile

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

Desktop Virtualization

- Citrix XenApp 6.5 – Published Desktops or HSD
- Session based desktops accessed using ICA

Hypervisor

VMware ESXi 5.1

XenApp Server Configuration

- Eight Microsoft Windows 2008 R2 Server VMs with 4vCPUs and 16 GB of memory per server
- Microsoft Windows Desktop Experience – Not Enabled

Server Specifications

- Cisco UCS B200 M3 with Dual 8-cCore Intel Xeon E5-2690 processors @ 2.90 GHz and 384GB RAM (24 x 16GB DIMMS @ 1666MHz)
- Cisco UCS VIC 1240 Virtual Interface Card- 4x10Gb
- Cisco VM-FEX was not used for this testing but including it here as a recommended configuration

Workload Profile: Cisco Knowledge Worker (ver3.3)

- Microsoft Office 2010 Applications
- Internet Explorer
- Adobe Acrobat 9
- Cisco Jabber for Windows (Version 9.1.3)
- 30sec Flash Video

Storage

NetApp FAS 3170 with PAM2 module (512 GB of cache)

Data Collection/Test Tool

- Workload Generation - 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 160 users were well within 5sec for all applications and meet the success criteria defined at beginning of this document.

Table 5 Response Times for 160 HSD users on Cisco UCS B200M3 with XA 6.5/ESX5.1/ICA/NetApp Profile

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 160 HSD users on Cisco UCS B200M3
Cisco Jabber for Windows (Version 9.1.3)	5s	0.6s
Outlook'10	5s	1.6s
Word'10	5s	0.9s
Excel'10	5s	0.8s
PowerPoint'10	5s	0.8s
Internet Explorer	5s	0.7s
Acrobat 10	5s	0.8s

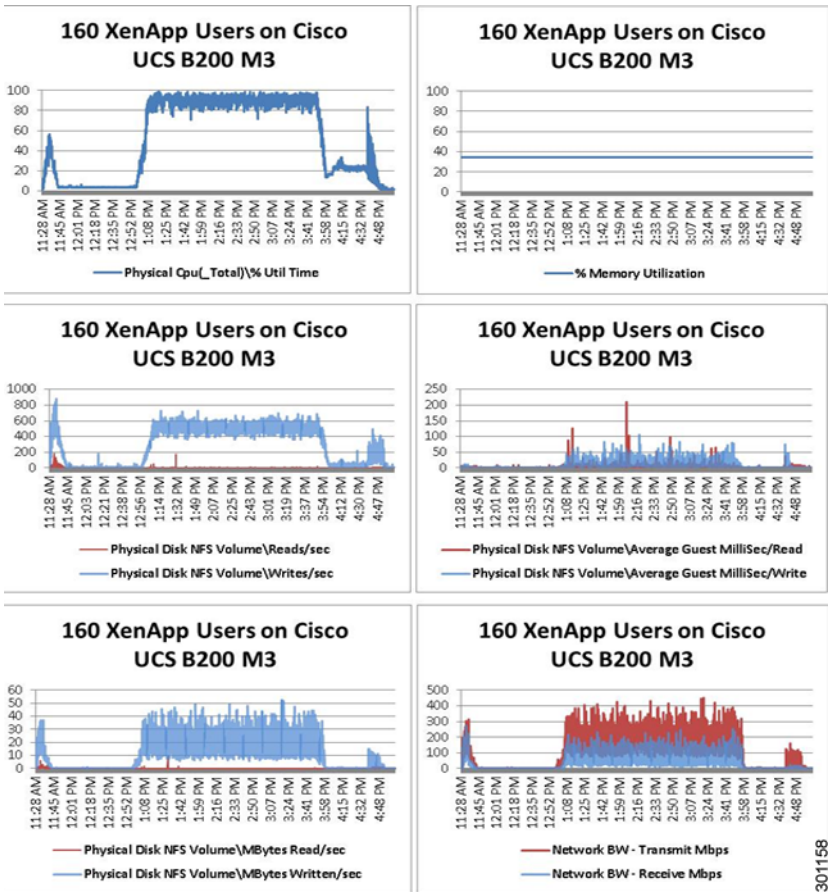
Server Performance Server Performance

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 8 XenApp servers supporting 160 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. This chart confirms that we are CPU bound for this profile as stated earlier. CPU Utilization is at 90% during steady state workload phase which represents all 160 users using their shared desktop sessions spread across eight XenApp server VMs.

The second chart is the memory utilization chart showing the memory allocation relative to the total memory deployed and this utilization is ~35%. Note that the utilization is low as the server was deployed with 384GB for this test though it is not needed. A memory configuration of 144GB of memory would have been sufficient for this workload with 16GB per XenApp Server (8 x 16GB =128GB) and including memory required for ESXi and other overhead. However, a performance optimized configuration on Cisco UCS B200M3 would require either 128GB or 192GB to be deployed. Deploying with 192GB of memory provides a good buffer to comfortably support 160 users with a Knowledge Worker workload. 128GB could also be sufficient for a lighter workload or if memory over-subscription is acceptable. However, assuming a 192GB configuration with no over-subscription, 10% buffer and memory for ESXi and virtualization overhead, still allows for ~1GB+ of memory for user session. Using 128GB with similar assumptions provides for ~680MB of memory per user session. For planning purposes, Citrix recommends using 500MB per user for a normal workload for calculating the memory needs of the deployment. Enterprises are still advised to assess the needs of their users through pilots and other in-house validation to ensure the accuracy of the estimation used in planning. This can be done through real time monitoring of the user base or by using a workload that closely matches an Enterprise user's desktop use in production.

Figure 2 *Performance Charts for 160 HSD users on Cisco UCS B200M3 with XA6.5/ESX5.1/ICA/NetApp Profile*



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 with 8 XenApp servers supporting 160 users, with each user running the Cisco KW+ workload. The read and write I/O load profile is as expected for a HSD environment with low write I/O and negligible read I/O – both read and write I/O for a HSD deployment will be significantly lower than an equivalent HVD deployment. The average I/O latency is <5msec – the average I/O latency should be < 20ms. A few peaks in the I/O latency are seen but the average is well within acceptable range and no impact to user experience was seen. The I/O Bandwidth data chart shows the network bandwidth utilization associated with the storage traffic – note that this is in Mbytes/sec. The peak and average I/O performance data shown in the charts above are also summarized in the table below.

Table 6 *I/O Performance during Steady State, Login and Logout of 160 user sessions*

Storage I/O	Steady State	Login	Logout
Read-Avg	1.73	23.78	0.31
Read-Peak	172.61	184.97	13.81
Write-Avg	509.34	394.04	81.02
Write-Peak	728.74	880.68	493.52
Read-Latency-Avg.	4.61	2.22	0.59
Read-Latency-Peak	209.74	6.28	35.66

Storage I/O	Steady State	Login	Logout
Write-Latency-Avg.	16.46	2.87	2.32
Write-Latency-Peak	104.72	12.05	75.17

The last chart shows the network bandwidth utilization which is a combination of the storage traffic and all other traffic sent and received by the XenApp server VM and can be a starting point for estimating the bandwidth needs of a HSD deployment in the data center. Note that due to the reduced storage traffic in a HSD environment, the overall network bandwidth generated is also less.

In summary, Citrix XenApp provides a scalable alternative to hosted virtual desktops with lower storage and memory requirement while scaling to a higher number of users per Cisco UCS server. However, it is important to understand that XenApp HSD delivery model is targeted for a different type of user than hosted virtual desktops. HSD is better suited for Task Worker or Light Knowledge worker type users. Enterprises should leverage both delivery models as needed to meet the needs of the different types of users.

HVD Scalability on Cisco UCS B200M3 with Citrix PVS write cache on local SSD

Cisco Virtual Workspace (VXI) Smart Solution brings technology advancements on Cisco UCS server to reduce the Total Cost of Ownership (TCO) associated with virtual desktop deployments. A significant impact to the virtual desktop TCO comes from storage and to address this, various storage optimization technologies like Citrix Intellicache and Atlantis ILIO has been developed to reduce the I/O load on the storage array. Another approach taken for reducing storage costs is to use a tiered storage architecture. One of the options using this approach is to use less expensive and therefore less reliable storage options, to front end the storage array, particularly for non-persistent users. This approach can be implemented by using local disks on the server, particularly the solid state drives, to meet some of the I/O demands. Using SSDs is best suited for a non-persistent desktop deployment since it has less stringent storage requirements.

Local SSD drives on Cisco UCS servers can be leveraged to reduce the storage array needs by serving the I/O locally. Multiple SSD drives can be deployed to further reduce the I/O going to the backend storage array. Cisco UCS B-series servers supports a maximum of 2 SSD drives today while the Cisco UCS C-series can support up to 24 drives depending on the model. In the Cisco Virtual Workspace system, testing was done to understand the scalability limits of a deployment using local SSD drives on a Cisco UCS B200M3 server. A non-persistent deployment using PVS streamed desktops were used in this testing.

Based on the testing done across the end-to-end Cisco Virtual Workspace system, Cisco UCS B200M3 can support 180 PVS based desktops with the PVS write cache for these desktops deployed on local SSD drives. Results indicate that we are CPU bound for this profile.

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 180 desktops on a Cisco UCS B200 M3 server with SSDs for its write cache.

Summary of Test Results

Using the above deployment profile, 180 desktops can be supported on a Cisco UCS B200M3 with the following performance metrics.

- Average CPU Utilization = ~90% (Steady state)

- Average Memory Utilization = ~30% (Installed memory was far higher than needed)
- Average IO Latency <15ms
- Average Steady State IOPS to SSD drives = ~1000 Writes
- Login Peak and Average Write IOPS = ~3500 and ~2500 Writes respectively
- Application Response times <2.5sec

Test Profile

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

Desktop Virtualization

PVS 6.1 based streamed desktops

Hypervisor

VMware ESXi 5.0 U1

Virtual Desktop Configuration

- Windows 7 32b desktops with 1.5G of RAM, 24G disk and 1vCPU per desktop
- Non-persistent desktop
- Write cache for desktop use is maintained in VM's 24G disk

Server Specifications

- Cisco UCS B200 M3 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
- 2 x 300G SSD drives in Raid 0 configuration

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

- Microsoft Office 2010 Applications
- Internet Explorer
- Adobe Acrobat9
- 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
- End user response times measured using Scapa TPP as outlined earlier in this document
- Data is captured and graphed for Login, Workload and Logout stages

Application Response Times

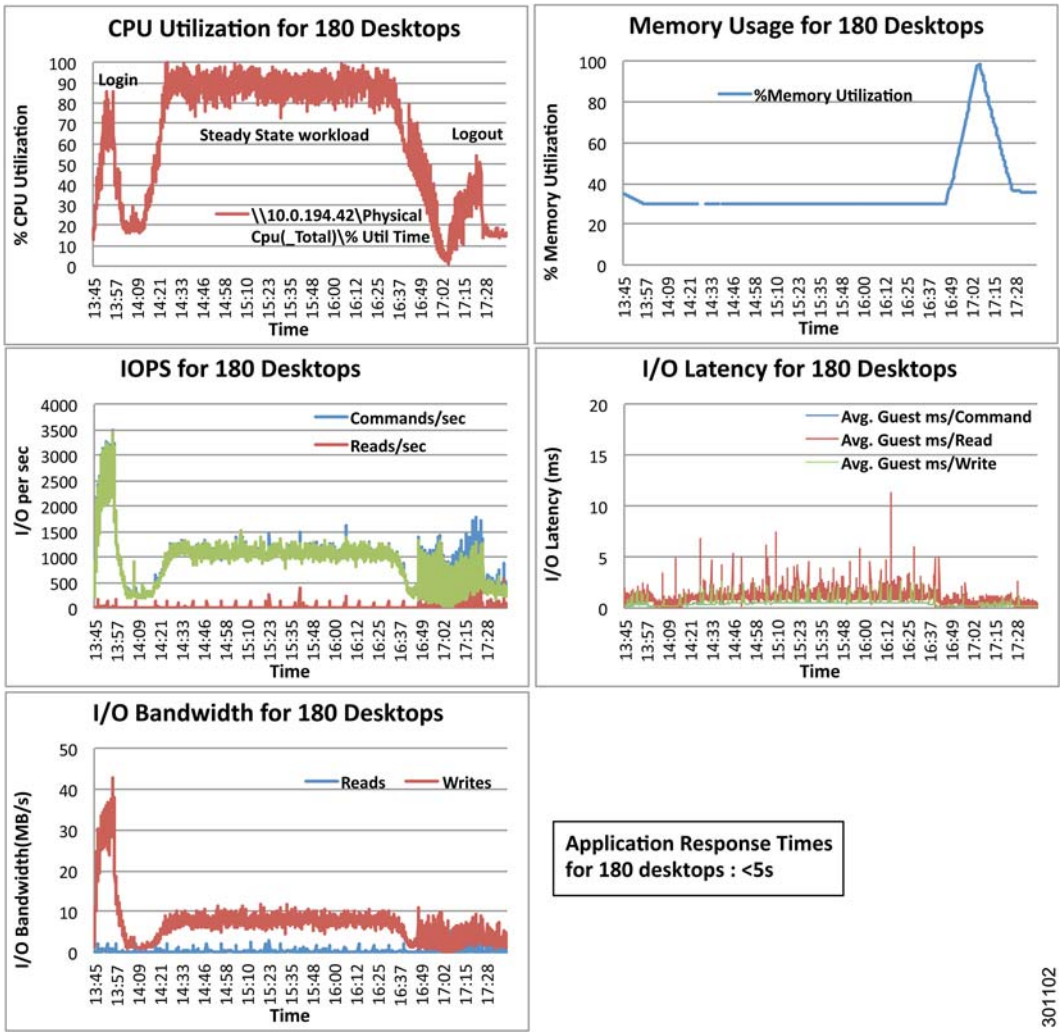
The table below shows that the response time experienced by 180 desktops were well within 5sec for all applications and meet the success criteria defined at beginning of this document.

Table 7 Response Times for 180 desktops on Cisco UCS B200M3 with PVS6.1/ESX5.0U1/ICA/NetApp Profile

Applications	Maximum Acceptable Startup Times(Success Criteria)	Average Startup Times Measured for 180 desktops on UCS B200M3
Cisco Unified Personal Communicator 8.5.1 in deskphone mode	5s	1.2s
Outlook'10	5s	2.2s
Word'10	5s	0.67s
Excel'10	5s	0.66s
PowerPoint'10	5s	0.71s
Internet Explorer	5s	0.71s
Acrobat	5s	0.66s

Performance Charts

Figure 3 Performance Charts for 180 desktops on Cisco UCS B200M3 with PVS6.1/ESX5.0U1/ICA/NetApp Profile



The first chart shows the CPU utilization measured using resxtop with a 5s polling interval during the Login, Workload and Logout stages of 180 desktops on a Cisco UCS B200 M3 server. This chart confirms that we are CPU bound for this profile with CPU utilization near 90%. CPU Utilization of 90% during steady state workload phase represents all 180 users using their desktop.

Memory utilization, on the other hand is around 30% during steady state. 384G of memory was deployed on the server but the data in the above graph show that only 30-35% of the total memory is needed for this deployment model.

The next few charts show the storage performance data. The IOPS chart shows the steady state IOPS are close to a 1000 Write IOPS, all of which is served by the local SSD.

In summary, SSD drives on Cisco UCS servers can be used to handle the PVS write cache in a PVS based virtual desktop deployment. This will send the Write I/O to the local SSD drives and offload the storage array from the more expensive Write I/Os.

XenDesktop 5 (XD5) on XenServer 5.6 FP1

This section covers the results of the single server scalability testing on a UCS B250 M2 with XenDesktop 5 on XenServer 5.6 FP1 running Windows 7 32b desktops and using NetApp for storage. Based on the testing done with Cisco KW+ workload, 110 HVDs can be supported on a UCS B250 M2 for this workload profile.

Test Environment and Setup

- XenDesktop 5 (using Provisioning Services) on XenServer 5.6 FP1
- HVD Profile:
  - Windows 7 32b with 1.5G of memory and 24G of disk space
  - 1 vCPU, Pooled 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 192G of memory - Two Six Core Intel Xeon (EP) 5680 processors @ 3.33 GHz
- Storage: NFS to NetApp FAS 3170 with PAM2 (512G of cache)
- All response times are measured using Scapa TPP as outlined above
- Data presented below is for login, workload and logout phases

Application Response Times

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

Table 8 Application Response Times for XD5 on XS 5.6 FP1 using a UCS B250 M2 and NetApp FAS 3170

Applications	Success Criteria for Maximum Acceptable Startup Times	Average Startup Times Measured (sec)
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	1.4s
Outlook	10s	2.1s
Excel	5s	82s
PowerPoint	5s	54s
Acrobat	5s	50s
Internet Explorer	5s	3.1s
Word	10s	7.0s

Summary of Test Results

For the deployment profile detailed in the Test Environment and Setup section above, 110HVDs can be supported on a Cisco UCS B250 M2 with the following performance metrics:

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = ~90%
- Storage
  - IOPS = Peak READ IOPS of <1800
  - IO Latency < 5ms
- Network Bandwidth Utilization = Peak BW utilization is <250Mbps during workload start

Figure 4 CPU Utilization on a UCS B250 M2 for Windows 7 32b/XD5/XS5.6 FP1 using NetApp FAS 3170

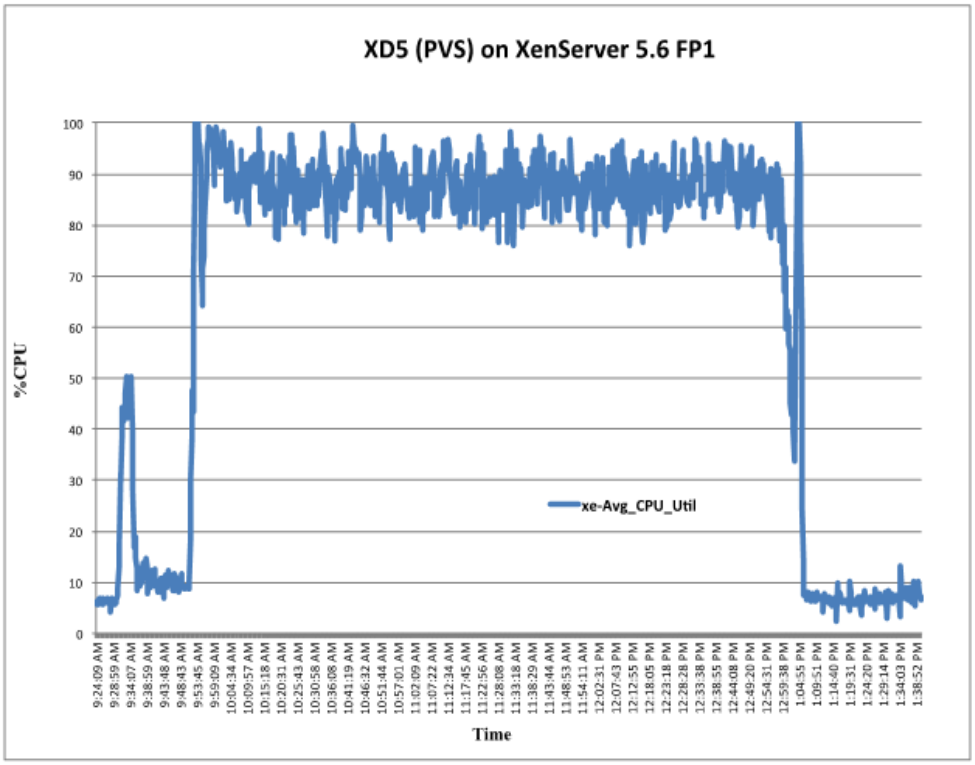


Figure 5 Memory Utilization on a UCS B250 M2 for Windows 7 32b/XD5/XS5.6 FP1 using NetApp FAS 3170

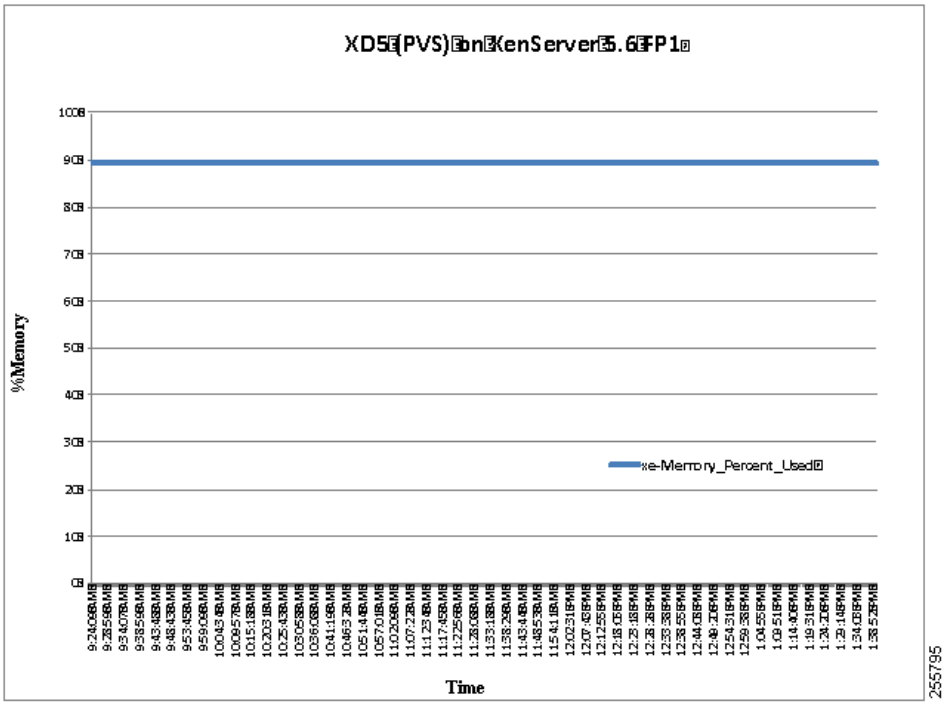


Figure 6 IOPS for Windows 7 32b/XD5/XS5.6 FP1 using NetApp FAS 3170

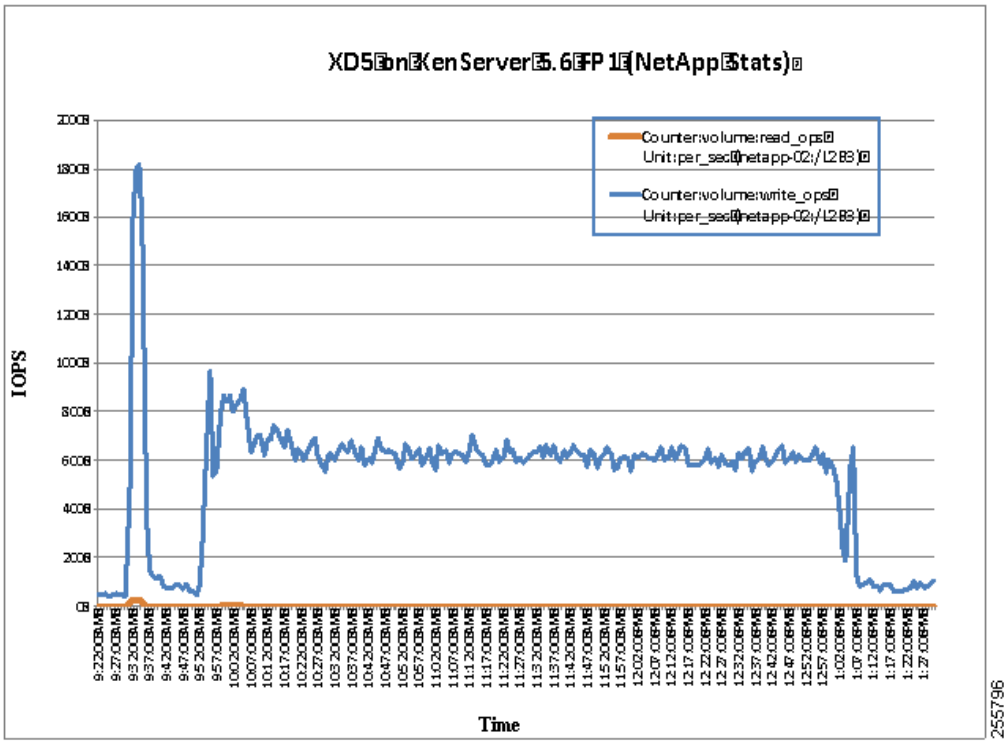


Figure 7 IO Latency for Windows 7 32b/XD5/XS5.6 FP1 using NetApp FAS 3170

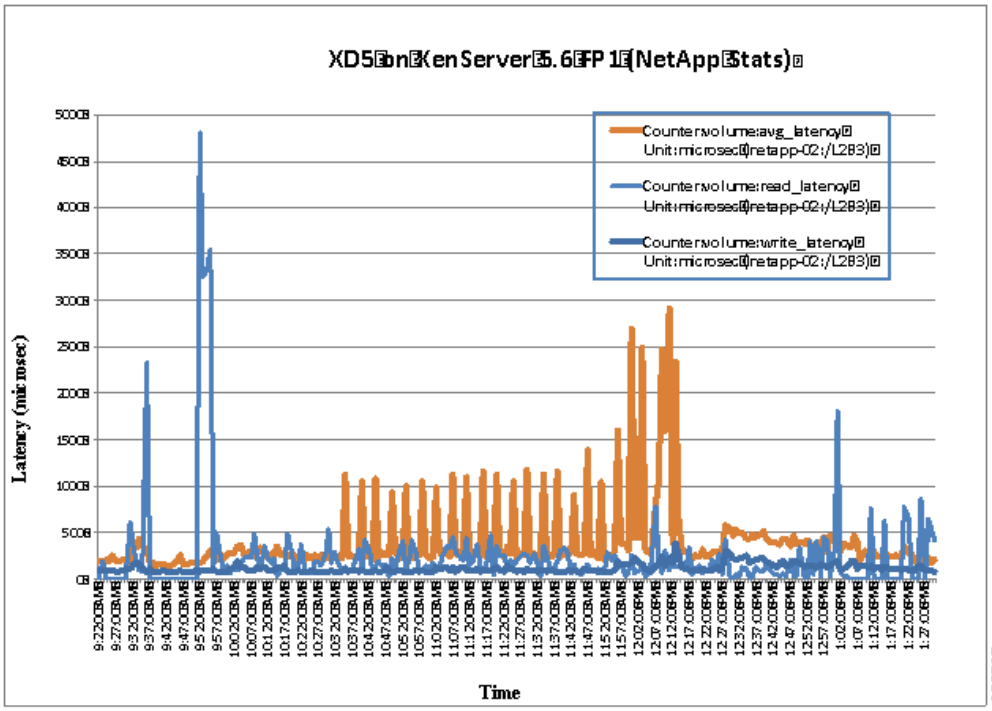
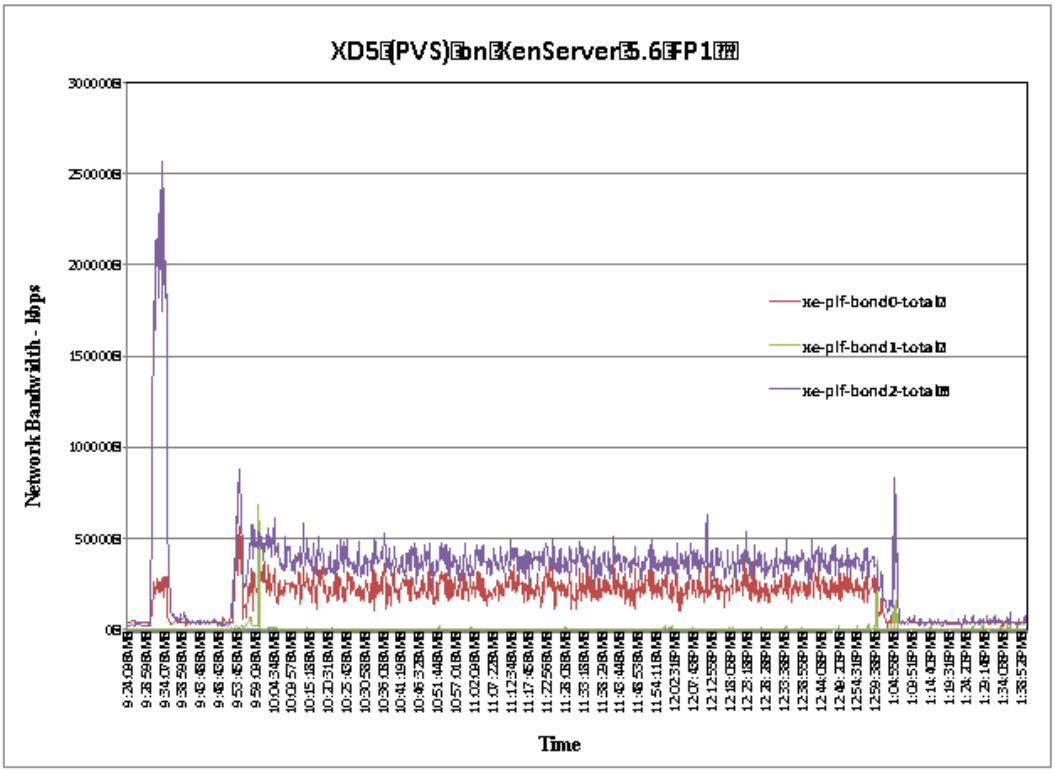


Figure 8 Network Bandwidth Utilization for Windows 7 32b/XD5/XS5.6 FP1 using NetApp FAS 3170



XenDesktop 5 (XD5) on ESXi

This section covers the results of the single server scalability testing on a UCS B250 M2 with XenDesktop 5 and ESXi running W7 32b desktops and using NetApp for storage. Based on the testing done with Cisco KW+ workload, 80 HVDs can be supported on a UCS B250 M2 with this profile.

Test Environment and Setup

- XenDesktop 5 (no Provisioning Server using Machine Creation Services) on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G of memory and 24G of disk space
  - 1 vCPU, Persistent Pooled desktop
- Workload Profile: Cisco KW+ (Cisco Unified Personal Communicator, 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: 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 5sec except for the NetApp view of IO statistics
- All response times are measured using Scapa TPP as outlined above
- Data presented below is for login, workload and logout phases
- For this profile, data is captured and graphed for the workload phase

Application Response Times

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

Table 9 Application Response Times for XD5 on ESXi on UCS B250M2 with NetApp

Applications	Success Criteria for Maximum Acceptable Startup Times	Average Startup Times Measured (sec)
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	3.6
Outlook	10s	2s
Excel	5s	6s
PowerPoint	5s	4s
Acrobat	5s	4s
Internet Explorer	5s	3s
Word	10s	7s

Summary of Test Results

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

- Average CPU Utilization = 90% (Steady state)
- Average Memory Utilization = 65%
- Storage
  - IOPS = Peak READ IOPS of 6500 seen during workload start
  - IO Bandwidth = Peak Read BW of 900MBps seen during workload start
  - IO Latency < 15ms
- Network Bandwidth Utilization = Peak BW utilization of over 1Gbps during workload start

Figure 9 CPU Utilization on a UCS B250M2 for W7-32b/XD5/ESXi 4.1 using NetApp

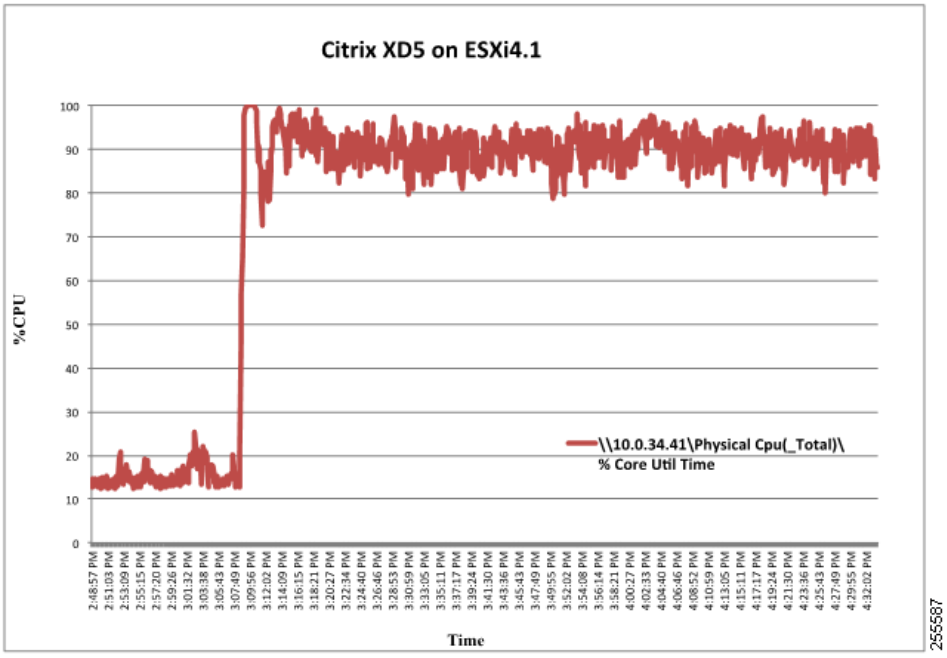


Figure 10 Memory Utilization on UCS B250M2 for W7-32b/XD5/ESXi 4.1 using NetApp

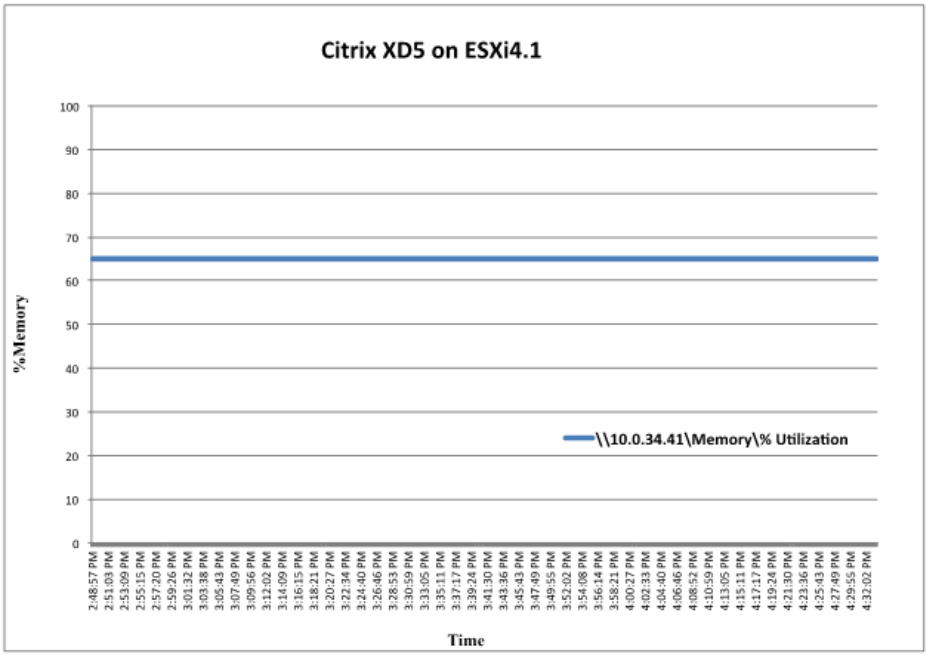


Figure 11 IOPS for W7-32b/XD5/ESXi 4.1 on UCS B250M2 using NetApp

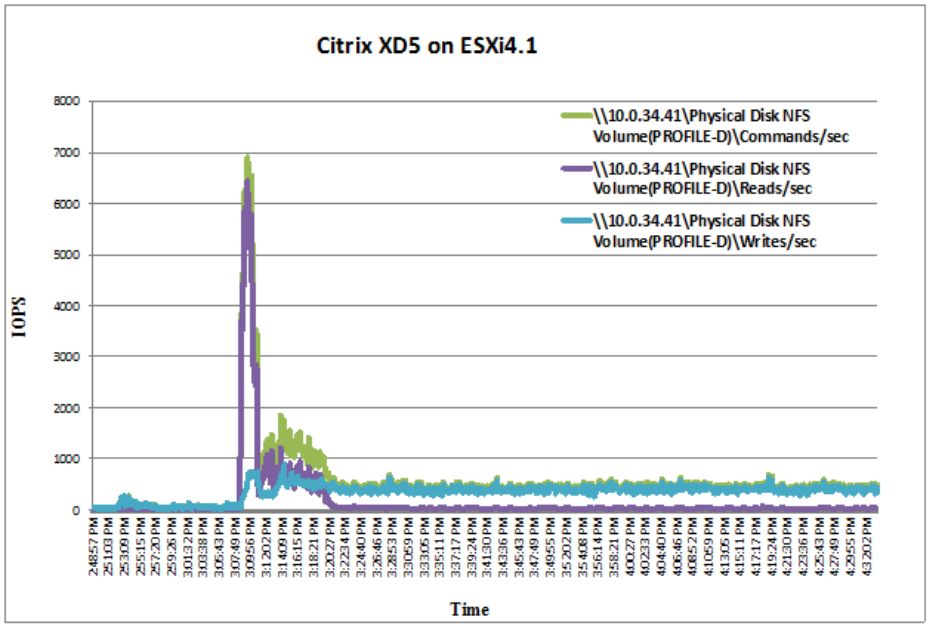


Figure 12 IO Latency for W7-32b/XD5/ESXi 4.1 on UCS B250M2 using NetApp

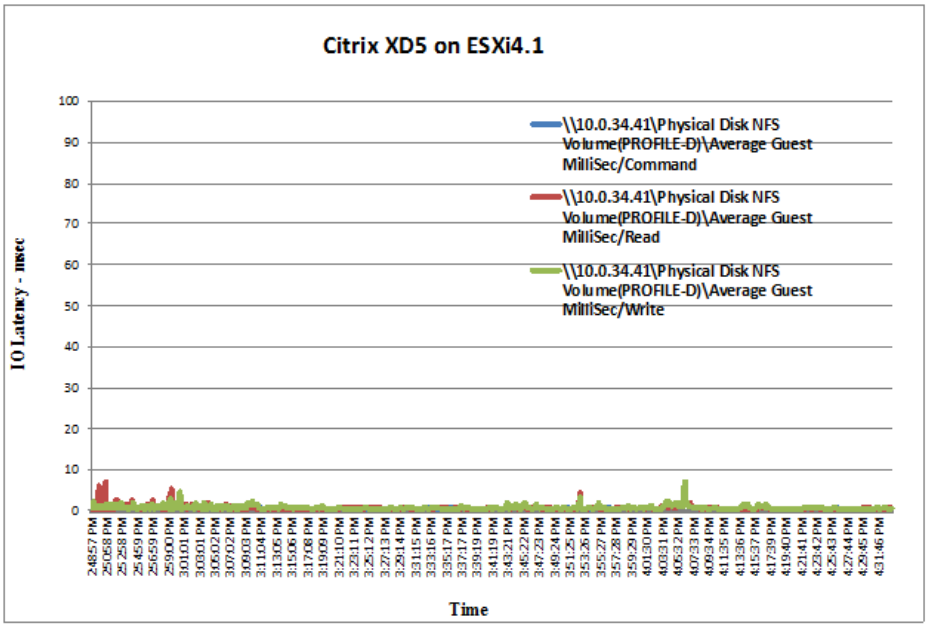


Figure 13 IO Bandwidth for W7-32b/XD5/ESXi 4.1 on UCS B250M2 using NetApp

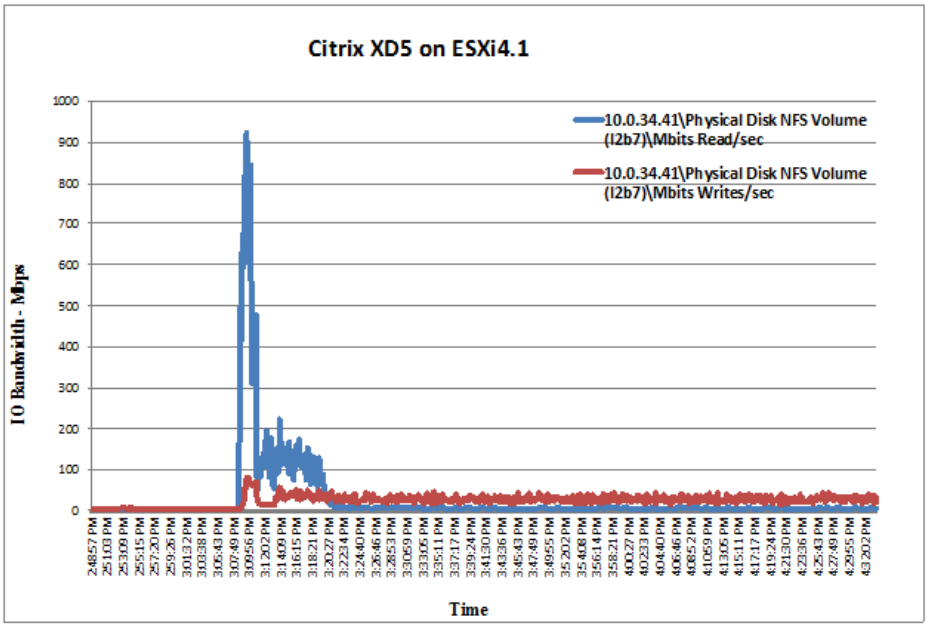


Figure 14 IO statistics from NetApp FAS 3170

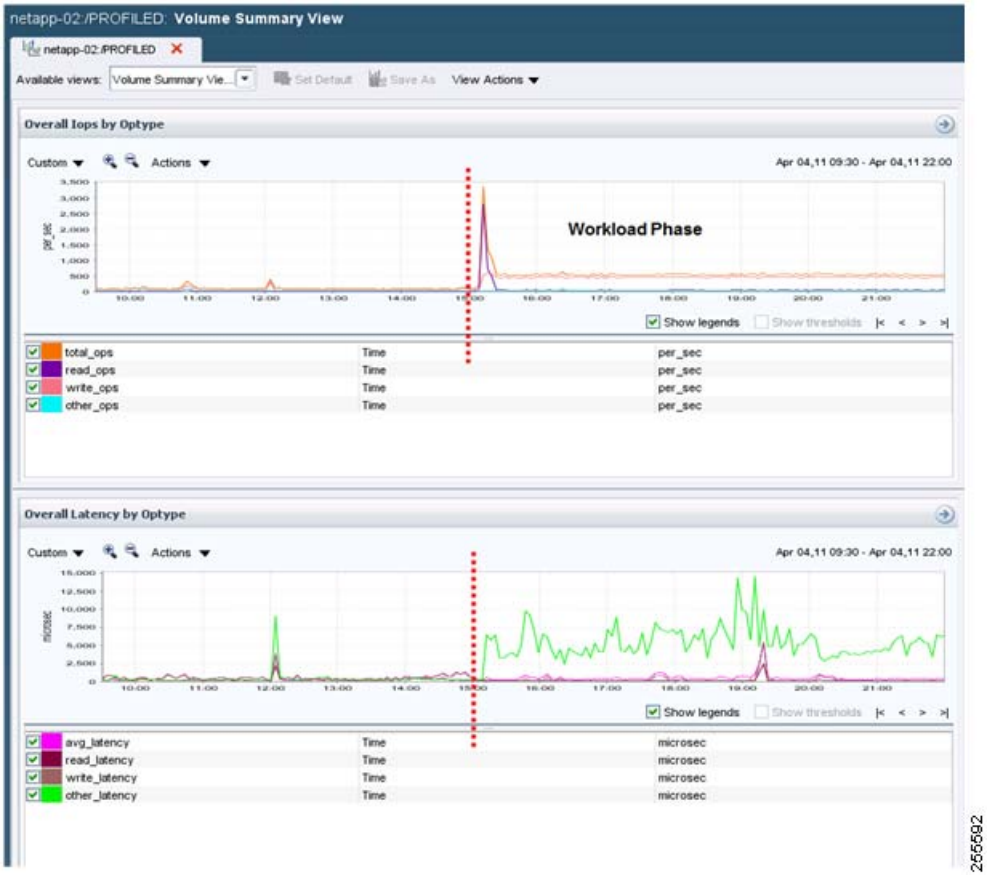
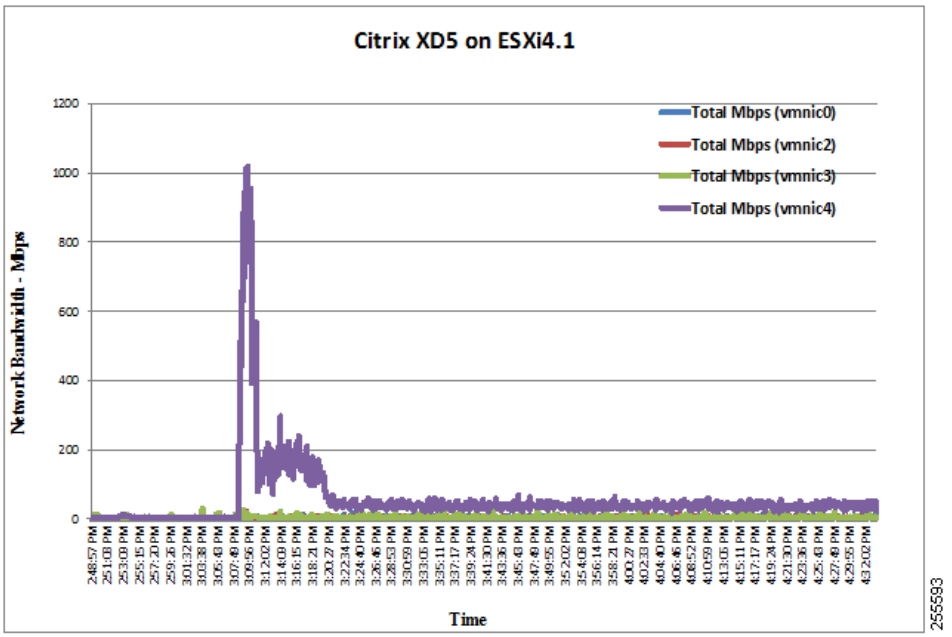


Figure 15 Network Bandwidth for W7-32b/XD5/ESXi 4.1 on UCS B250M2 using NetApp



XenDesktop5/ESXi4.1U1/ICA/B250M2 Profile –New CPU Utilization 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 XenDesktop5 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 response time bound for this profile.

Test Profile

Desktop Virtualization

- XenDesktop 5 using Machine Creation Services
- Connection Protocol – ICA
- Pooled Static

Hypervisor

VMware ESXi 4.1 U1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Non-persistent – Refresh after logoff disabled for ease of testing

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 = ~67% (Steady state)
- Average Memory Utilization = 80%
- Application Response times – Success Criteria met

Performance Charts

Figure 16 CPU Utilization Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile)

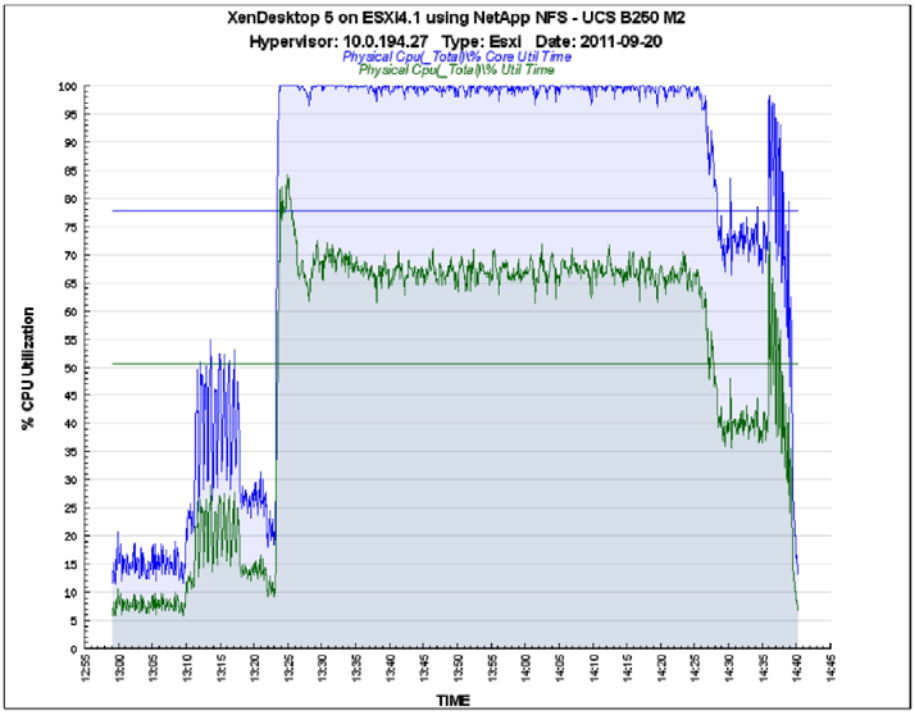


Figure 17 Memory Utilization Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile)

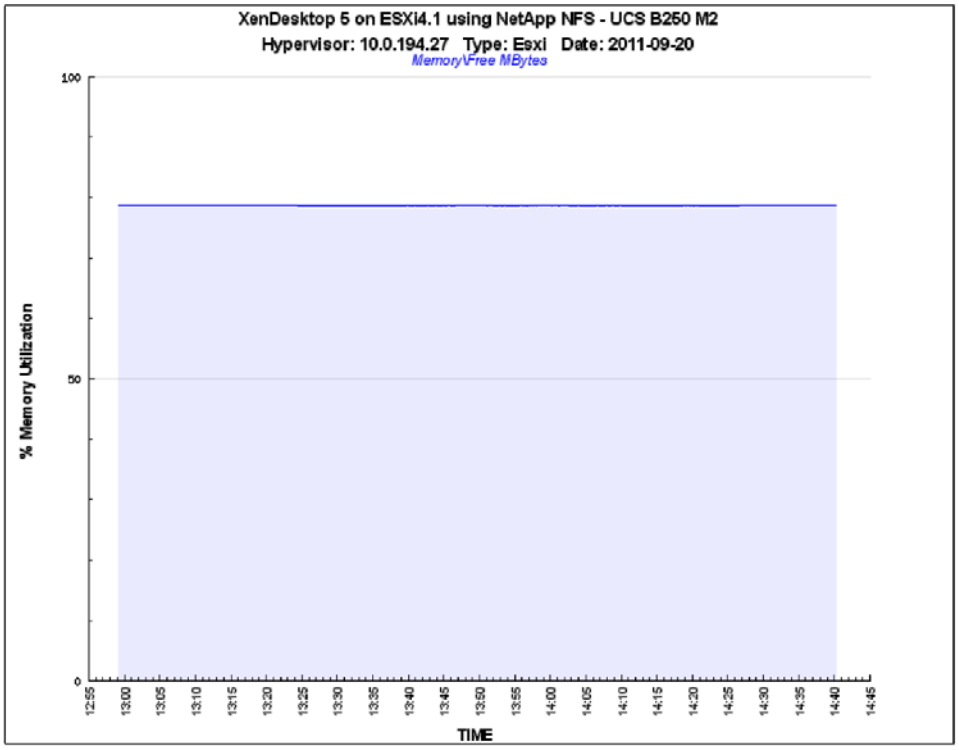


Figure 18 IOPS Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile)

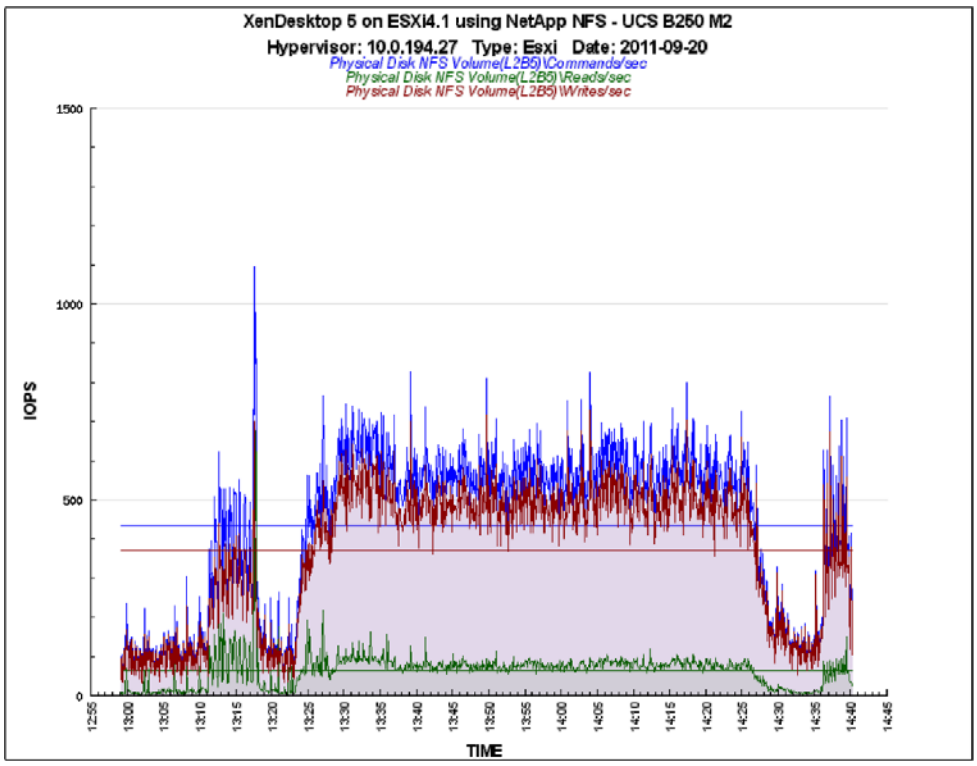


Figure 19 IO Latency Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile

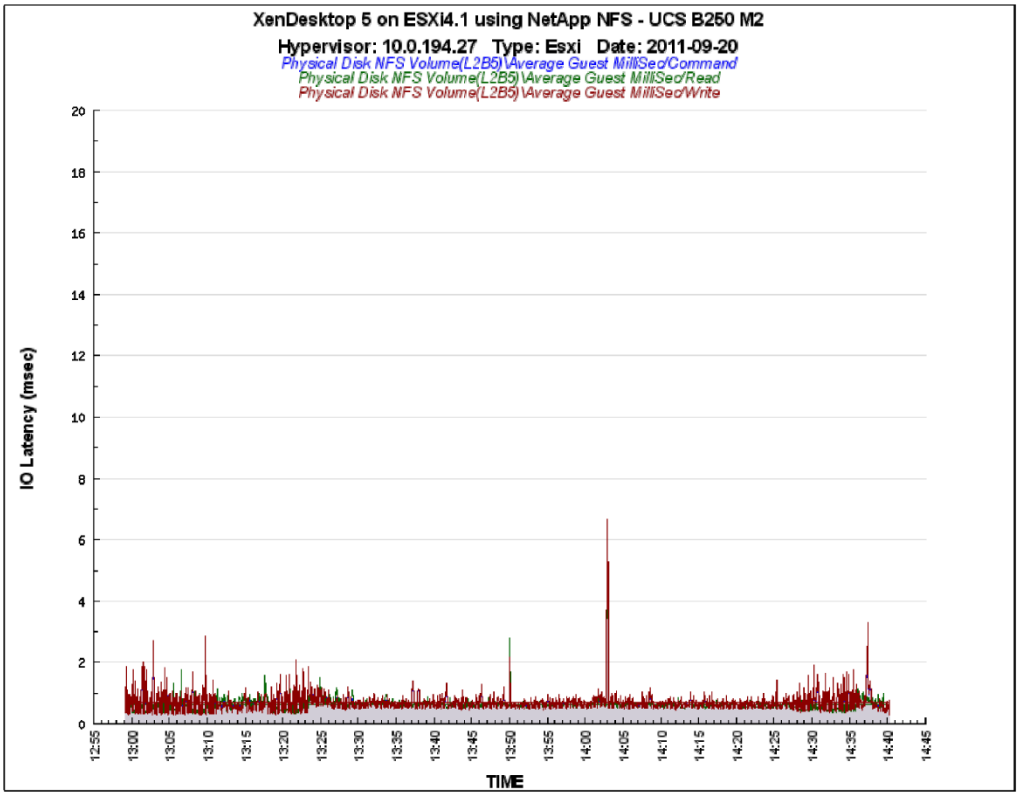


Figure 20 IO BW Utilization Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile

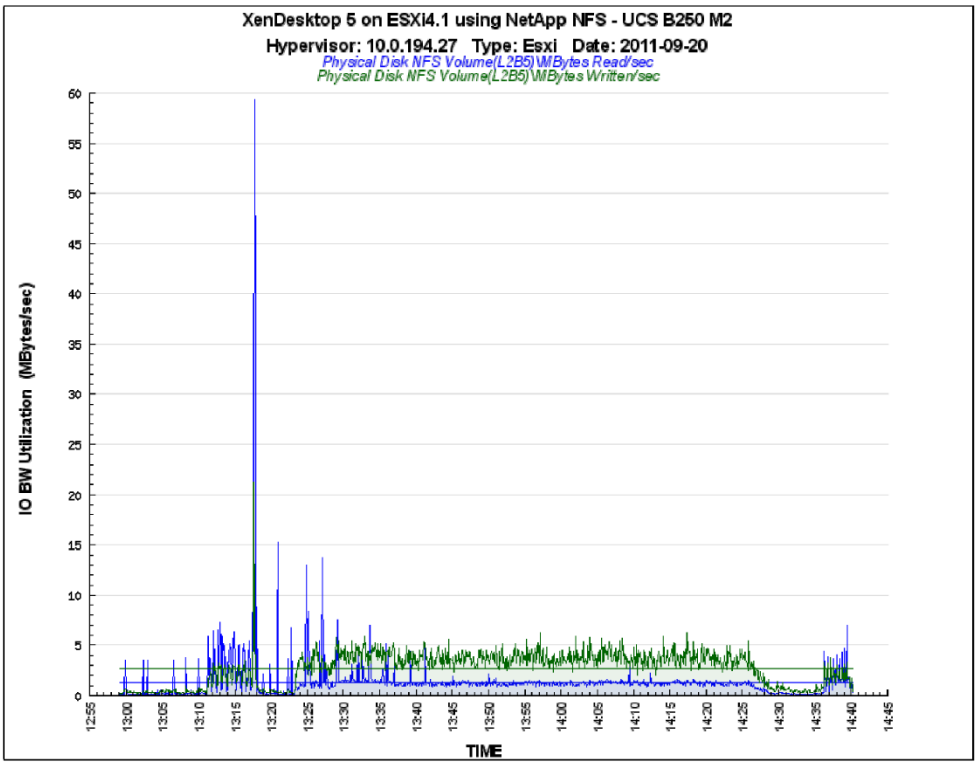
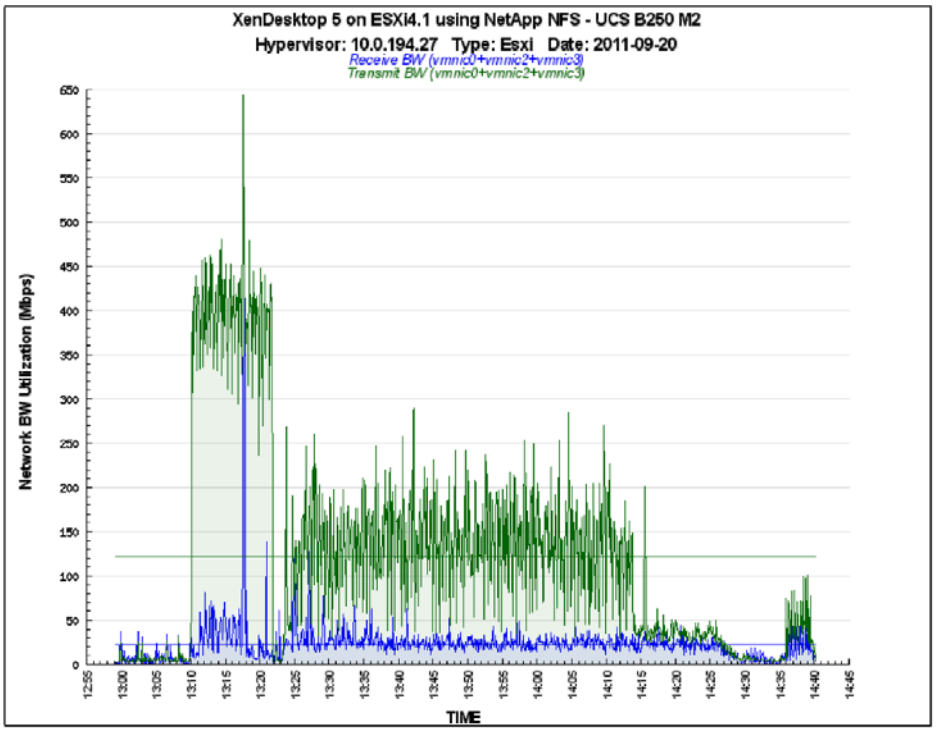


Figure 21 Network BW Utilization Chart for XenDesktop5/ESXi4.1U1/ICA/B250M2/NetApp Profile



Application Response Times

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

Table 10 Response Times for XenDesktop5.5/ESXi4.1U1/ICA/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	4.6s
Word	10s	8.5s
Excel	5s	1.2s
Powerpoint	5s	0.75s
Internet Explorer	5s	4.4s
Acrobat	5s	0.63s

XenDesktop5.5/ESXi5/ICA/B230M2 Profile - FlexPod

This section provides the detailed results of the single server scalability tests done for a UCS B230 M2 across a FlexPod infrastructure with Windows 7 32b desktops running on XenDesktop 5.5 and ESXi5. 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 CPU bound for this profile.

Test Profile

Desktop Virtualization

- XenDesktop 5.5 using Machine Creation Services
- Connection Protocol – ICA
- Pooled Static

Hypervisor

VMware ESXi5

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Non-persistent – Refresh after logoff disabled for ease of testing

Server Specifications

- Cisco UCS B230 M2
- Two Ten Core Intel Xeon E7-2870 @ 2.40GHz

- 256G RAM (32 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

- 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, 160 virtual desktops can be supported on a Cisco UCS B230 M2 with the following performance metrics.

- Average CPU Utilization = ~90% (Steady state)
- Average Memory Utilization = ~92%
- Application Response times – Success Criteria met

Performance Charts

Figure 22 CPU Utilization Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile)

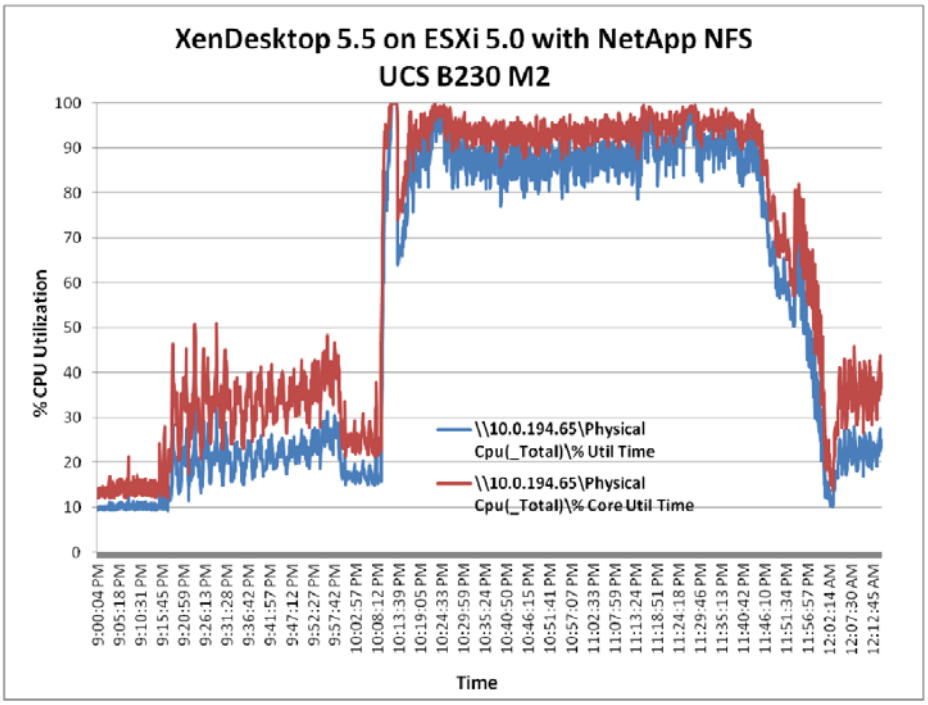


Figure 23 Memory Utilization Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile)

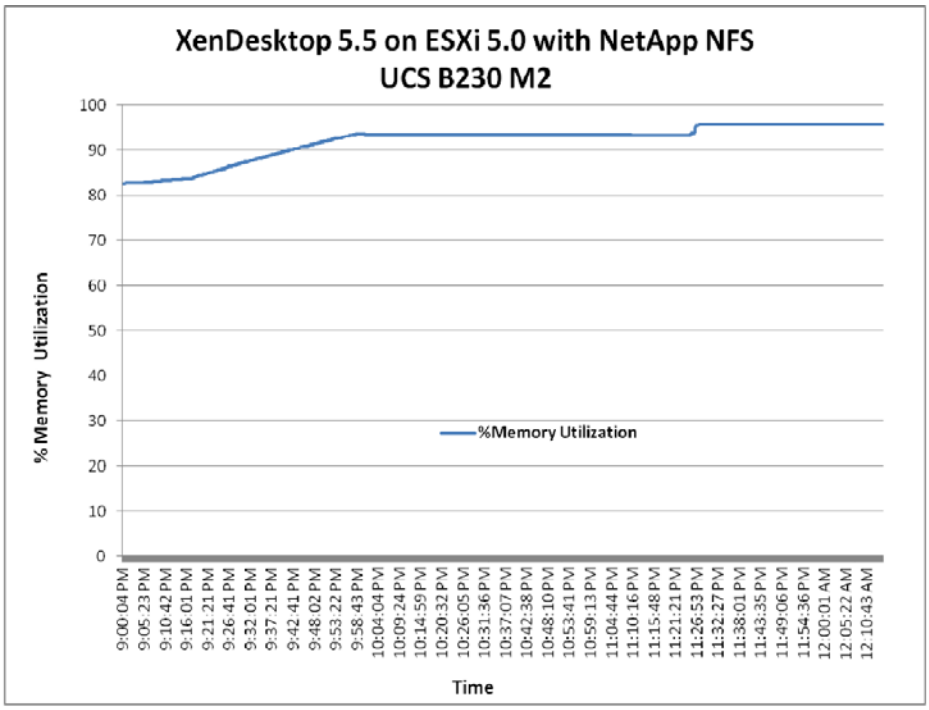


Figure 24 IOPS Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile)

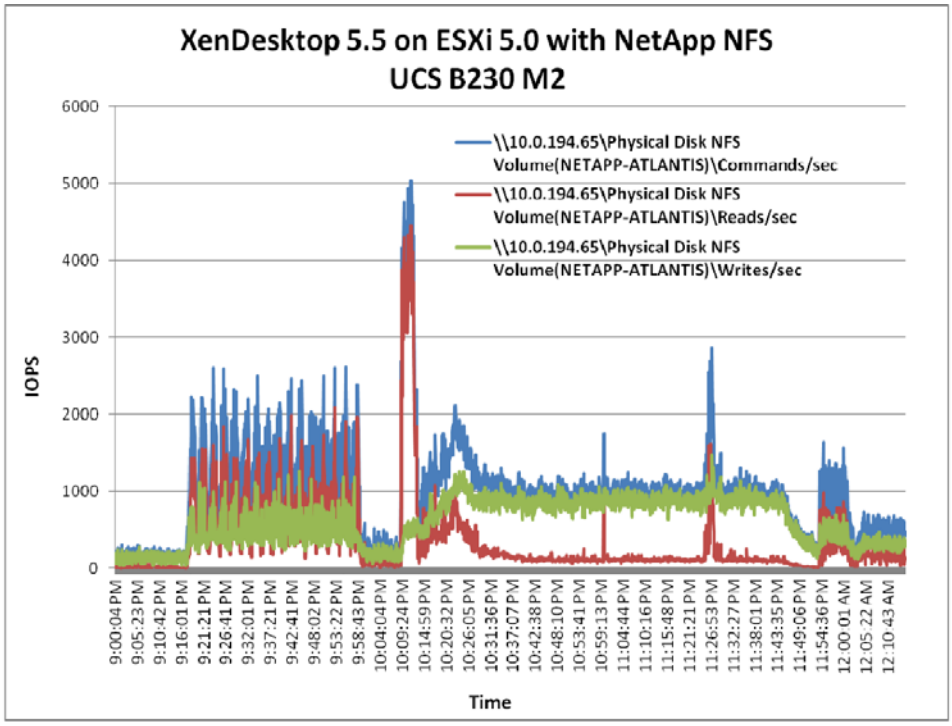


Figure 25 IO Latency Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile

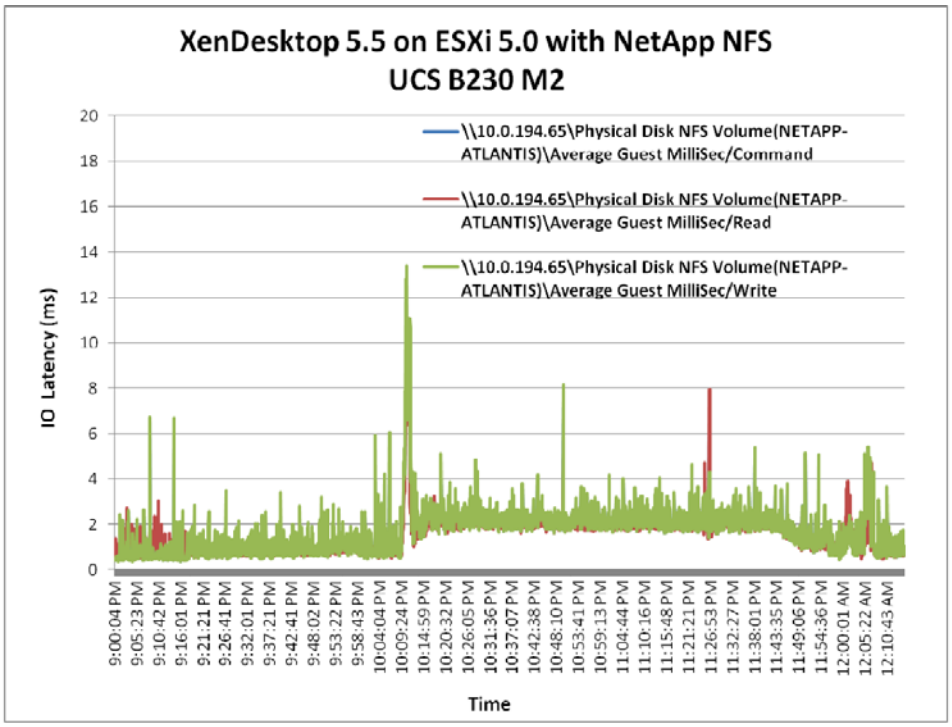


Figure 26 IO BW Utilization Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile

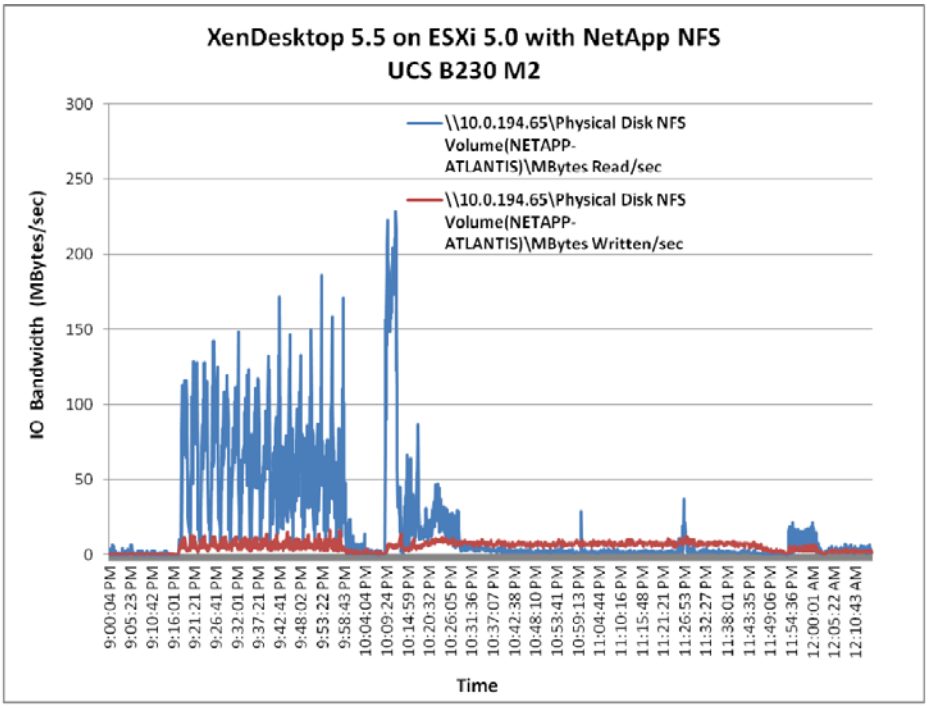
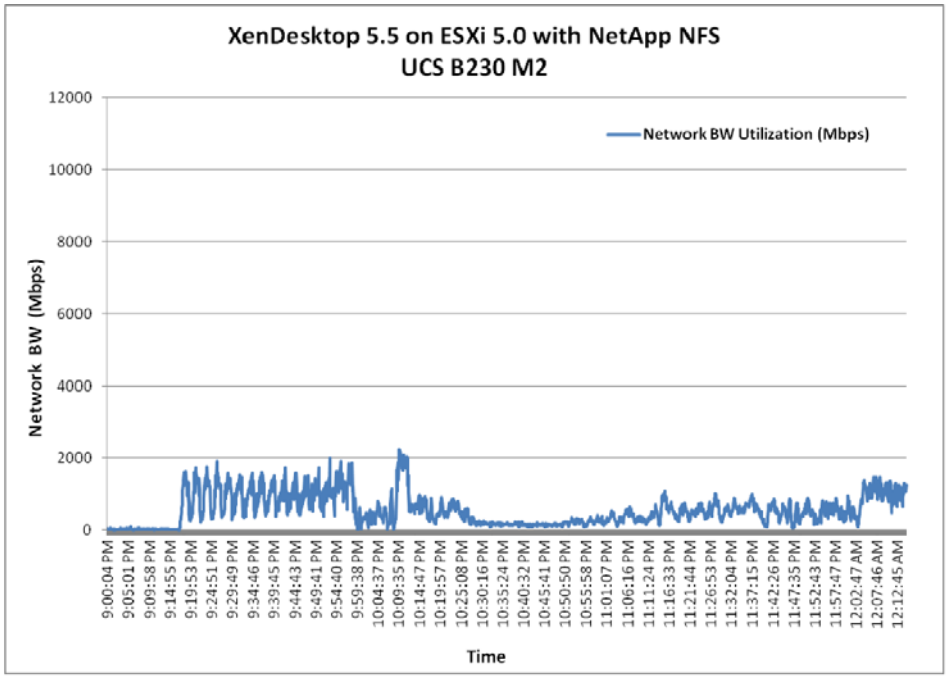


Figure 27 Network BW Utilization Chart for XenDesktop5.5/ESXi5/ICA/B230M2/NetApp Profile



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 XenDesktop5.5/ESXi5/ICA/B230M2/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	1.6s
Outlook	5s	2.6s
Word	5s	.83s
Excel	5s	1.1s
Powerpoint	5s	0.77s
Internet Explorer	5s	1.1s
Acrobat	5s	0.68s

XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M Profile - Static Memory

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 with 96G of memory and Windows 7 32b desktops running on XenDesktop 5.5 and Hyper-V 2008 R2 SP1. Results indicate that ~60 virtual desktops can be supported on a UCS B250M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are memory bound for this profile.

Test Profile

Desktop Virtualization

- XenDesktop 5.5 using Machine Creation Services
- Connection Protocol – ICA
- Pooled Static

Hypervisor

Microsoft Hyper-V 2008 R2 SP1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Non-persistent – Refresh after logoff disabled for ease of testing

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz
- 96G RAM (12 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
- Antivirus software was excluded in this profile due to issues seen

Storage

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

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Perfmon for statistics collection from hypervisor

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

Summary of Test Results

For the deployment profile detailed above, 60virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = ~30% (Steady state)
- Average Memory Utilization = 98% but since this is with static memory, HyperV ensures every HVD has 1.5G of memory allocated to it and as such operating at high memory utilization is less of a concern
- Application Response times – Success Criteria met

Performance Charts

Figure 28 CPU Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile)

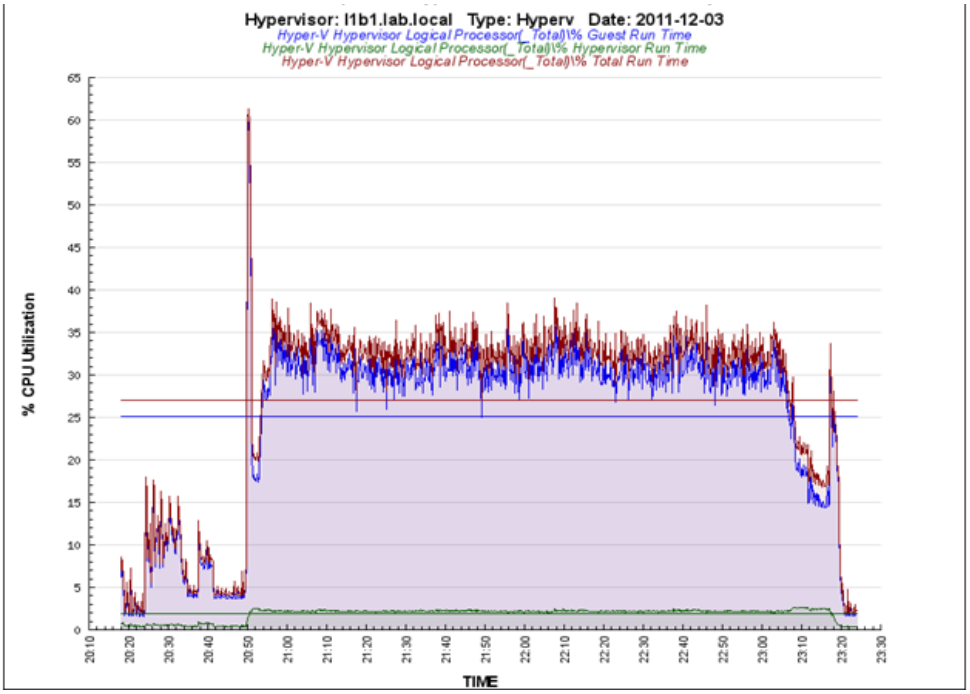


Figure 29 Memory Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/ NetApp Profile)

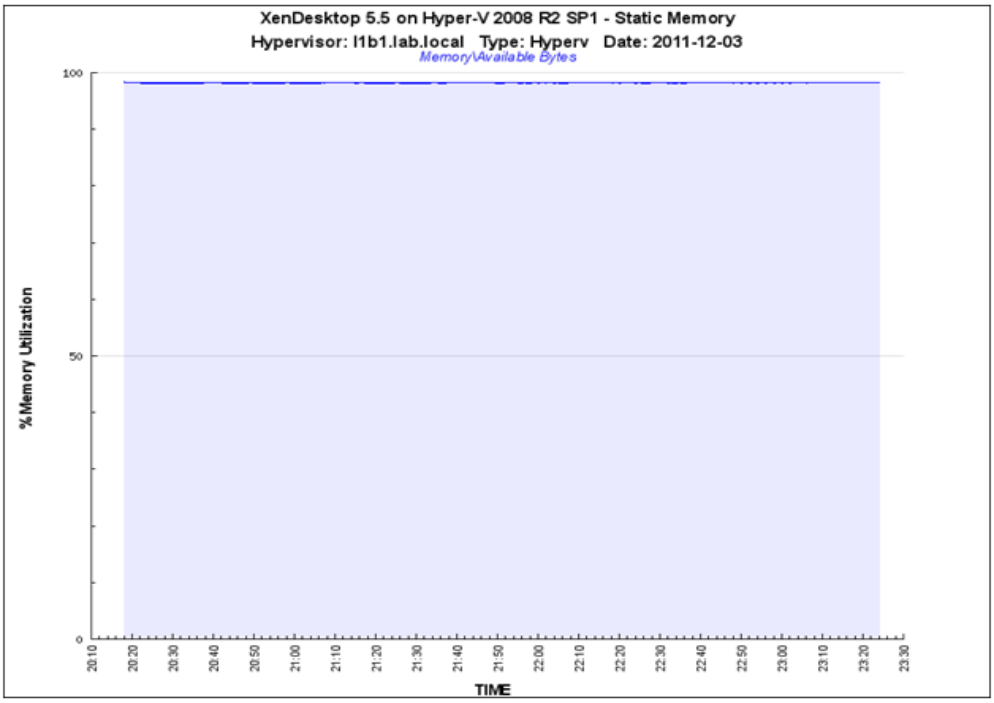


Figure 30 IOPS Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile

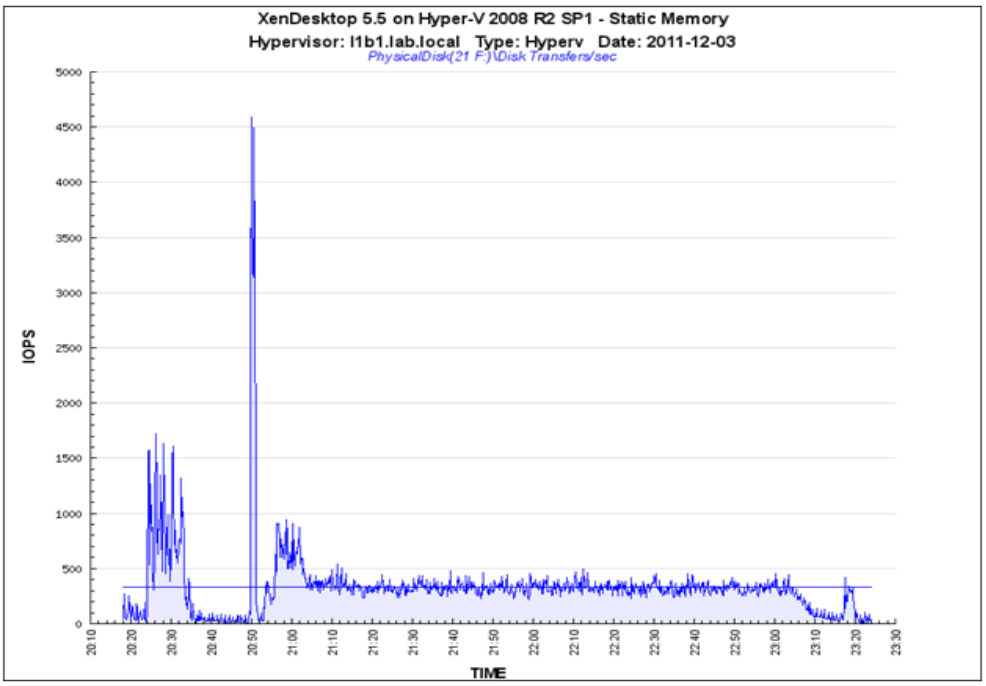
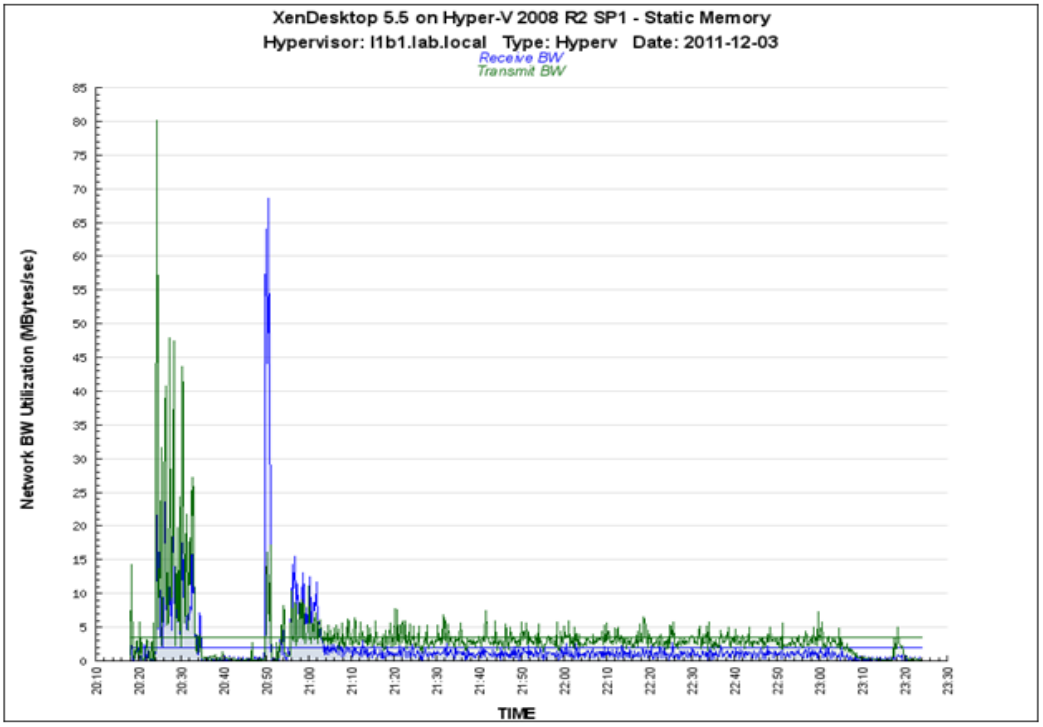


Figure 31 Network BW Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile



Application Response Times

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

document.

Table 12 Response Times for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp with Static Memory 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	3.2s
Outlook	5s	1.9s
Word	5s	0.62s
Excel	5s	0.67s
Powerpoint	5s	0.46s
Internet Explorer	5s	0.58s
Acrobat	5s	0.42s

XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M Profile – Dynamic Memory

This section provides the detailed results of the single server scalability tests done for a UCS B250 M2 with 96G of memory and Windows 7 32b desktops running on XenDesktop 5.5 and Hyper-V 2008 R2 SP1. Results indicate that ~100 virtual desktops can be supported on a UCS B250M2 based on testing done with a Cisco KW+ workload. Results also indicate that we are memory bound for this profile by using 96G of memory.

Test Profile

Desktop Virtualization

- XenDesktop 5.5 using Machine Creation Services
- Connection Protocol – ICA
- Pooled Static

Hypervisor

Microsoft Hyper-V 2008 R2 SP1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Non-persistent – Refresh after logoff disabled for ease of testing

Server Specifications

- Cisco UCS B250 M2
- Two Six Core Intel Xeon (EP) 5680) processors @ 3.33 GHz

- 96G 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
- Antivirus software was excluded in this profile due to issues seen

Storage

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

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Perfmon for statistics collection from hypervisor
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login, Workload and Logout phases

Summary of Test Results

For the deployment profile detailed above, 100 virtual desktops can be supported on a Cisco UCS B250 M2 with the following performance metrics.

- Average CPU Utilization = ~52% (Steady state)
- Average Memory Utilization = 91%
- Application Response times – Success Criteria met

Performance Charts

Figure 32 CPU Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile)

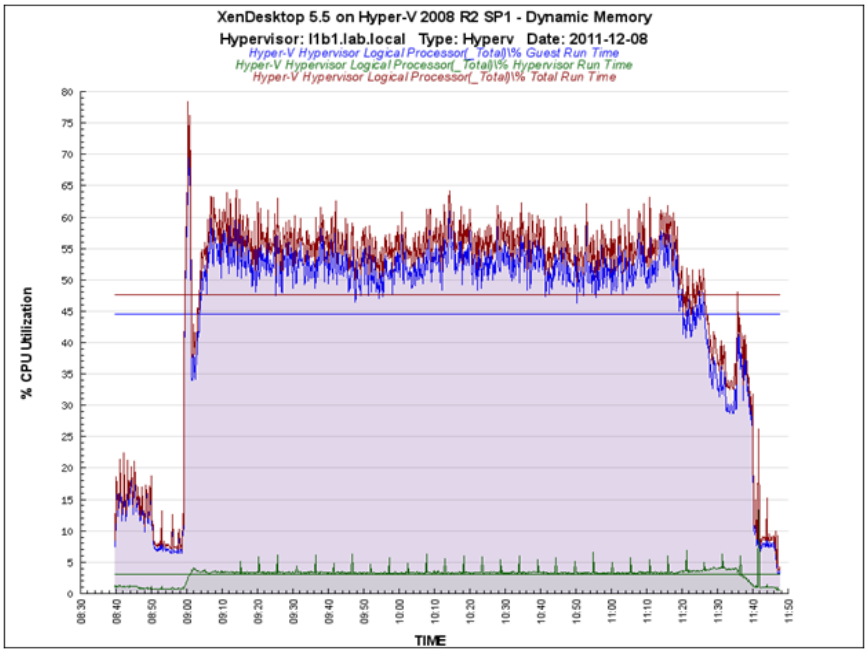


Figure 33 Memory Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile)

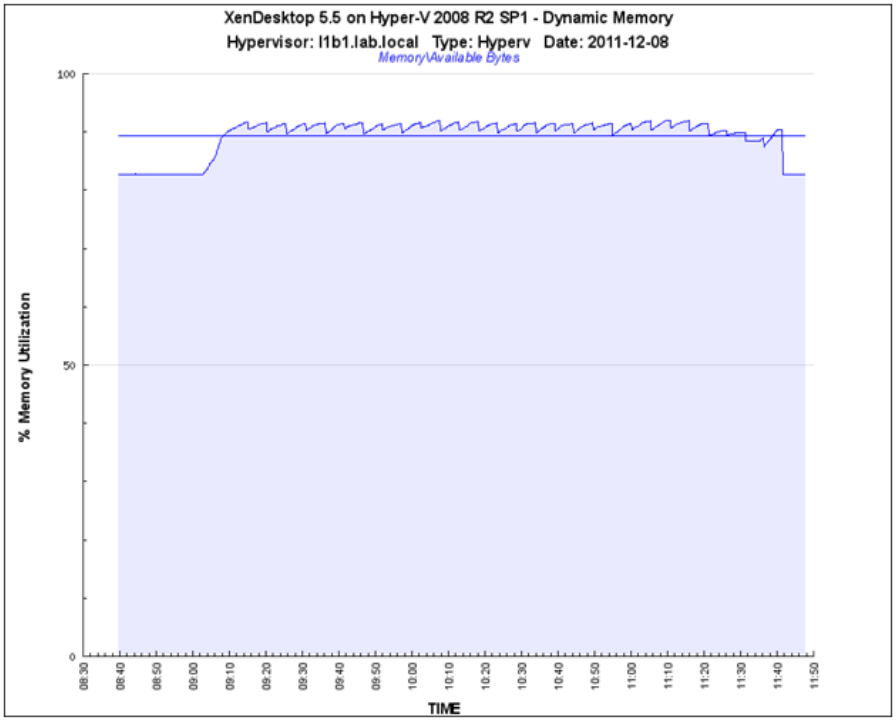


Figure 34 IOPS Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile)

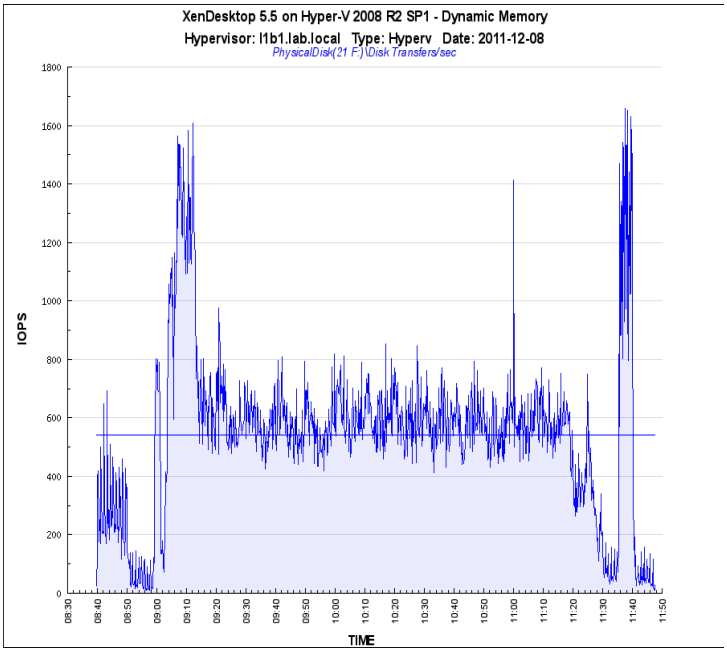
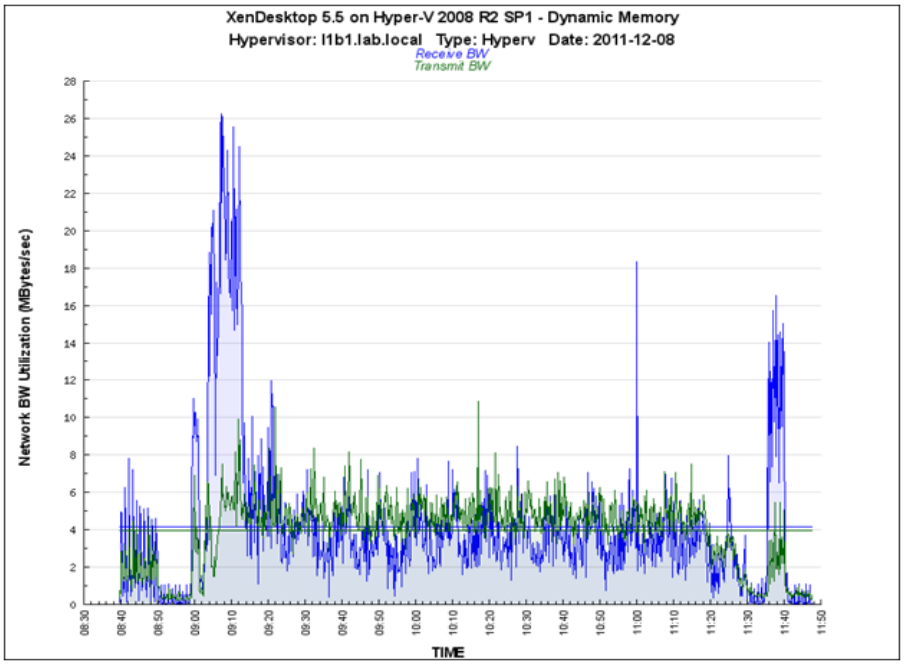


Figure 35 Network BW Utilization Chart for XenDesktop5.5/Hyper-V-2008R2-SP1/ICA/B250M2/NetApp Profile



Application Response Times

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

Table 13 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	1.6s
Outlook	5s	2.0s
Word	10s	0.65s
Excel	5s	0.78s
Powerpoint	5s	0.55s
Internet Explorer	5s	0.72s
Acrobat	5s	0.63s

Performance Impact of Storage Optimization using IntelliCache

The objective of this test profile is to characterize the impact of Citrix IntelliCache for deploying hosted virtual desktops on Cisco UCS B230 M2 using XenDesktop. IntelliCache uses the server's local SSD drives as its cache with NetApp serving as the back end storage array. Citrix IntelliCache is a storage optimization solution available on XenServer hypervisor that builds a cache based on IO traffic being sent and received from the backend storage array. Once the cache is built, all subsequent read and write IO requests from the virtual machines on the server are served from this local cache, thereby reducing the read and write IO requests to the backend array. As more servers are deployed with IntelliCache , IO load on the storage array is further reduced. IntelliCache is particularly beneficial in desktop virtualization deployments where IO performance requirements are driving the high cost of storage which is a critical factor in the adoption of desktop virtualization.

IO profile for desktop workloads can have a high degree of variability between peak and average IOPS. In desktop workloads, peaks in Read IOPS are often seen during bootup and login while the Write IO is relatively low and steady. However, once the user starts using the desktop, Read IOPS drops becomes relatively low and steady (no peaks). Peaks in Read IOPS may again be seen during logout or power down which is also when Write IOPS can peak. As multiple users are booting up, logging in or logging out, these peaks can get significantly high and the storage array will have to be sized to meet these peak IO requirements so that user experience is not impacted. However, using technologies like IntelliCache enables one to suppress these peaks and size the storage array for lower IO loads, leading to significant cost savings.

In the Cisco Virtual Workspace system, Citrix IntelliCache provided a 98%+ reduction in the IOPS going to the backend storage array based on the testing done with a Cisco KW+ workload.

To deploy Cisco UCS servers with IntelliCache in a new desktop virtualization deployment, it is important to understand the optimal server configuration, maximum number of virtual desktops per server (density), the percentage of IOPS that can be offload and ensure that the virtual desktop provides a good user experience.

With this objective in mind, the testing was done with the UCS server running at maximum density where the maximum density was determined by the success criteria defined earlier in the document. Citrix IntelliCache was enabled on a Cisco UCS B230M2 server with the virtual desktops and used the local SSD drives as its cache. For this test, 2x64G SSD drives in a RAID0 configuration was used. The desktops were deployed as pooled desktops on the backend NetApp storage array. To establish the maximum density, a load test was run that included logging on to Windows 7 virtual desktop and launching applications including Microsoft Office applications, Adobe Acrobat, internet browsing and

the Cisco Unified Personal Communicator (CUPC). The number of desktops running the load test was increased until one of the Success Criterion was reached. For this profile, with IntelliCache enabled, the test was stopped when the response times for launching applications reached a threshold of 5s at a maximum density 130 desktops. The test was then repeated without IntelliCache and showed that the response times were beyond the maximum response time of 5s. Therefore, without IntelliCache, maximum density possible on the server would less than 130 desktops. As such, IntelliCache not only provides IO offload but also improves the application response times, thereby improving the number of desktops that can be supported on a server.

Results Analysis Summary

- Testing done in the Cisco Virtual Workspace system shows that using Citrix IntelliCache significantly reduces the write and read IOPS to the backend storage array by as much as 98%+. In this testing, 130 desktops were deployed as pooled desktops on a Cisco UCS B230 M2 server using Machine Creation Services (MCS) with Citrix XenDesktop 5.5. IO performance charts from NetApp used as the backend storage array in the Cisco Virtual Workspace system is available in the Detailed Performance Results section of this document. See Table 14 below for more details.

Table 14 IOPS Optimization using Citrix IntelliCache

	IOPS seen on NetApp storage without IntelliCache	IOPS seen on NetApp storage with IntelliCache	IOPS Offloaded by Citrix IntelliCache
Average Read/Write IOPS (Workload Phase)	~600+	~1	~100%
Average Read/Write IOPS (Login Phase)	~200+	~3	~98%

- For the given workload, HVD density achieved on the Cisco UCS B230 M2 was higher by using IntelliCache than without it. With IntelliCache , 130 desktops could be supported based on the defined success criteria but without it, the response times were not within the success criteria for the same HVD density and so the number of HVD supported without IntelliCache would have to lowered.

Design & Deployment Considerations

- Newer Cisco UCS servers have much higher compute capacity (e.g. Dual Ten core processors) and improved cache designs that provide significant performance improvements at high server loads. These improvements have increased the overall virtual desktop densities supported on Cisco UCS servers and Citrix IntelliCache further improves this performance while reducing storage costs.
- IntelliCache is only available when using Machine Creation Services (MCS) – not available with Citrix Provisioning Services
- IntelliCache is currently supported only on local SSD drives. SSD is recommended for performance reasons. Enterprise grade SSD drives are available on Cisco UCS servers. Cisco UCS servers can support two SSD drives per server. UCS B230 M2 used in this testing can support either 2x64G or 2x100G SSD drives. For more information – see http://www.cisco.com/en/US/products/ps10280/prod_models_comparison.html

- Master image should be maintained on shared storage and MCS can be used to deploy desktops on the shared storage. IntelliCache will build the Read and Write Cache on the local SSD as users start launching and using their desktops.
- IntelliCache leverages the local SSD storage on the server to provide some of the benefits of local storage but without loss of centralized management and other operational benefits. Backend storage array is still required to house the master image and virtual desktops but without the need for expensive disks for meeting the IO performance needs of the workload.
- IntelliCache uses the SSD drivers for read and write caches. If the SSD caches gets exhausted, XenServer will fall back automatically to the back end storage. Maximize local SSD storage when possible to minimize fall back (e.g. 2x100G SSD drives on Cisco UCS servers). However, a pre-deployment evaluation in your environment can provide a more accurate estimation of the cache sizes and therefore the SSD storage needs. A small subset of users can be used in this evaluation to determine the read and write cache sizes after they've used their desktops for a period of time. The read cache should be the size of the shared master image. The individual write cache for each user should be the average for a larger group of users if a fairly representative subset of users were selected for this evaluation, Based on these two data points, the size of the cache can be determined as follows:

```
SSD Storage Capacity = Read Cache Size + (# of users x Avg. size of per-user Write
Cache)
```

Note Above should be the minimum capacity of the SSD drives used for IntelliCache.

- Thin provisioning must be enabled on the local SSD for IntelliCache which also will change local storage type from LVM to EXT3
- When using local SSD drives, the performance of the local storage becomes critical – some factors to consider are:
  - SSD performance starts to drop at higher utilization so IntelliCacheconsider leaving some headroom when sizing for IntelliCache in your deployment. Also, recommend monitoring servers running with IntelliCache as the SSD drives get closer to maximum capacity.
  - RAID 1 is typically used for local storage but RAID 0 provides significantly higher Write performance so you may want to consider using RAID 0. The risk of losing user customization data is minimized as data is backed up to shared storage array for persistent (dedicated) desktops – see below for more details. For non-persistent desktops, there is no data to be preserved and if the master image and desktops are maintained on shared storage, the data lost is any changes made during that session. This may be an acceptable tradeoff given the performance benefit - at least something for you to consider.
- IntelliCache is supported for non-persistent (pooled) desktops and maintains a local write cache for temporary info but the data is not written back to the backend storage array at log-off. Therefore, changes made to the desktop image, including applications installed or other user customizations are not preserved. This is to be expected since these are non-persistent desktops. For this reason, the backend storage array is used for reading the base image as the first VM powers up and logs on. However, once these blocks of data corresponding to the image are cached by IntelliCache, backend storage array is no longer accessed thus minimizing the Read and Write IO operations from the server to the array.
- IntelliCache is also supported for persistent (dedicated) desktops and works similar to the non-persistent case above. However, the temporary and user data is concurrently written to tw locations, to the local SSD write cache and also to the backend storage array resulting in Write IO from the server to the array. The Write IO benefit is less in this case but the Read IO benefit is still the same as in the non-persistent case.

- If the HVD density is high on a server, memory on XenServer Dom0 should be increased to 2940MB
- IntelliCache is a XenServer feature and is not supported on other Hypervisors
- XenMotion and High Availability is only available for persistent desktops. However the local cache file is not deleted from the host that the desktop was migrated from and the local cache on the new host is built as data is read.
- For more information on IntelliCache, see links below
  - How to use IntelliCache with XenDesktop: http://support.citrix.com/article/CTX129052
  - Improve XenDesktop TCO with XenServer IntelliCache: http://www.citrix.com/English/ps2/products/subfeature.asp?contentID=2317190
  - XenDesktop and local storage + IntelliCache: http://blogs.citrix.com/2011/06/22/xendesktop-and-local-storage-intellicache/

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, 130 virtual desktops can be supported on a Cisco UCS B230 M2 with the following performance metrics.

Server Metrics

- Average CPU Utilization = 80% (Steady state)
- Average Memory Utilization = ~80%
- Application Response times – Success Criteria met
- Average IO latency <20ms

Test Profile

Desktop Virtualization

- XenDesktop 5.5 using MCS
- Connection Protocol – ICA
- Pooled Desktops - Static

Hypervisor

Citrix XenServer 6.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Pooled desktop (reboot on Logout disabled for testing purposes)

Server Specifications

- Cisco UCS B230 M2
- 2 x Ten Core Intel Xeon E7-2870 processors @ 2.40GHz
- 256G RAM (32 x 8G DIMMS @1066 MHz)
- 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 desk phone 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
- IOSTAT with a polling interval of ~20s was used along with NetApp Operations Manager for IO data
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured and graphed for Login, Workload & Logout phases

In the next section, we look at the overall performance data with and without IntelliCache enabled.

Performance Charts

Figure 36 Performance Charts with Citrix IntelliCache for 130 desktops (XD5.5/XS6/ICA/B230M2/NetApp)

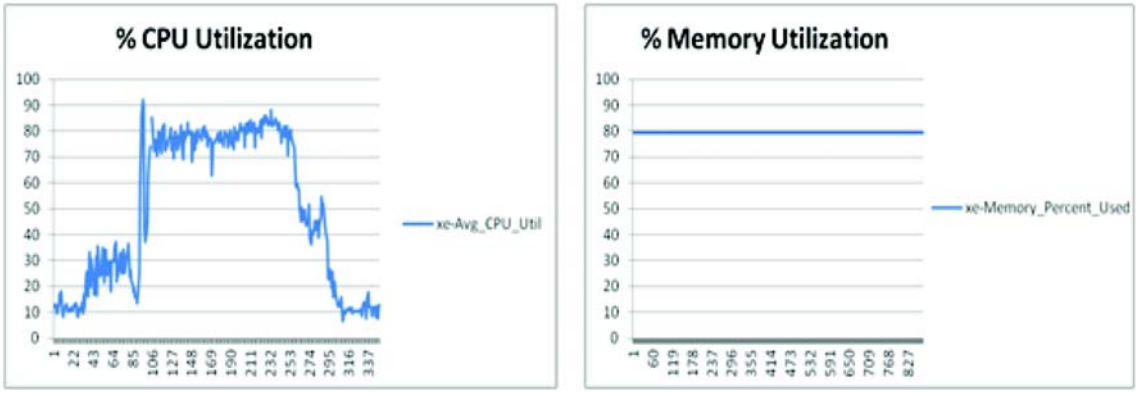


Figure 37 IO Charts for Baseline without IntelliCache for 130 desktops (XD5.5/XS6/ICA/B230M2/NetApp)

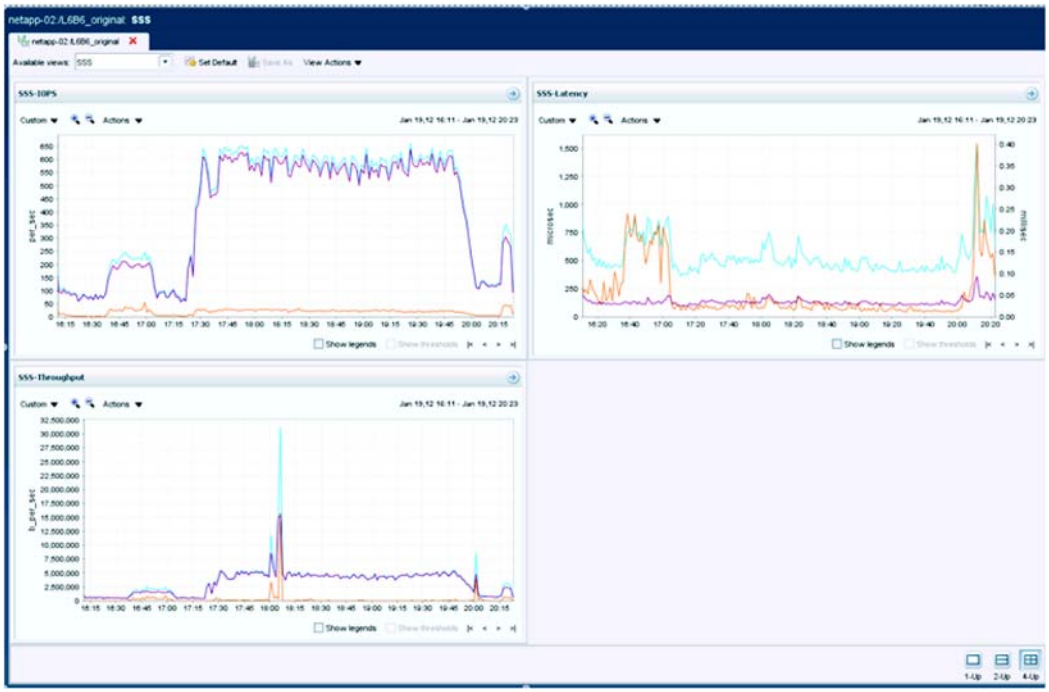
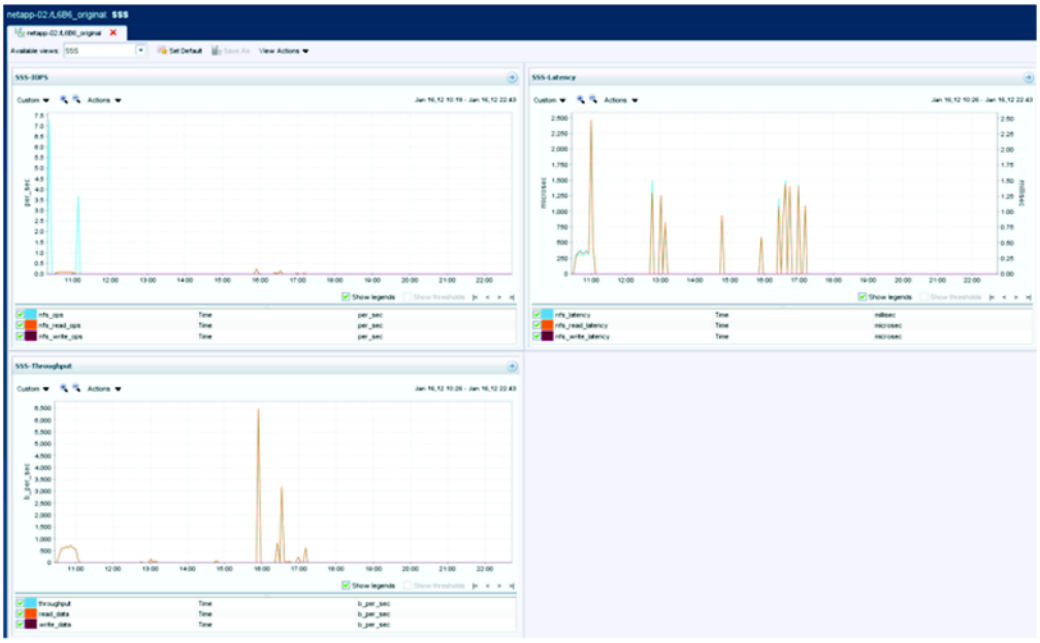


Figure 38 IO Charts with Citrix IntelliCache for 130 desktops (XD5.5/XS6/ICA/B230M2/Local SSD+NetApp)



Application Response Times

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

Table 15 Response Times for for Citrix IntelliCache Profile (XD5.5/XS6/ICA/B230M2/NetApp)

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test (With IC/Without IC)
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	1.5s/1.4s
Outlook	5s	5s/5.5s
Word	5s	1.8s/2.0s
Excel	5s	1.9s/2.1s
PowerPoint	5s	1.6s/1.7s
Internet Explorer	5s	2.6s/3.2s
Acrobat	5s	1.0s/1.0s

HVD Scalability for XenDesktop5.5/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 XenDesktop 5.5 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 and with significant transparent page sharing in effect, the memory utilization decreases

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 = ~80% with ~17% transparent page sharing
- Application Response times – Success Criteria met
- Average IO Latency <20ms (Actual = <10ms)

Test Profile

Desktop Virtualization

- Citrix XenDesktop 5.5
- Connection Protocol – ICA
- Pooled Static Desktops using Machine Creation Services (MCS)

Hypervisor

VMware ESXi 5.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Non-Persistent desktop but reboot on logout disabled for ease of testing

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 39 Performance Charts for XenDesktop5.5/ESXi5/ICA/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 16 Response Times for XenDesktop5.5/ESXi5/ICA/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	1.7s
Outlook	5s	2.8s
Word	5s	0.8s
Excel	5s	1.1s
PowerPoint	5s	0.8s

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Internet Explorer	5s	1.1s
Acrobat	5s	0.7s

Performance Baseline for Citrix XenDesktop 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 ~120 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

- Citrix XenDesktop 5.5
- Connection Protocol – ICA
- Pooled Desktops - Static

Hypervisor

VMware ESXi 4.1U1

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per 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

- 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 & Logout phases

Summary of Test Results

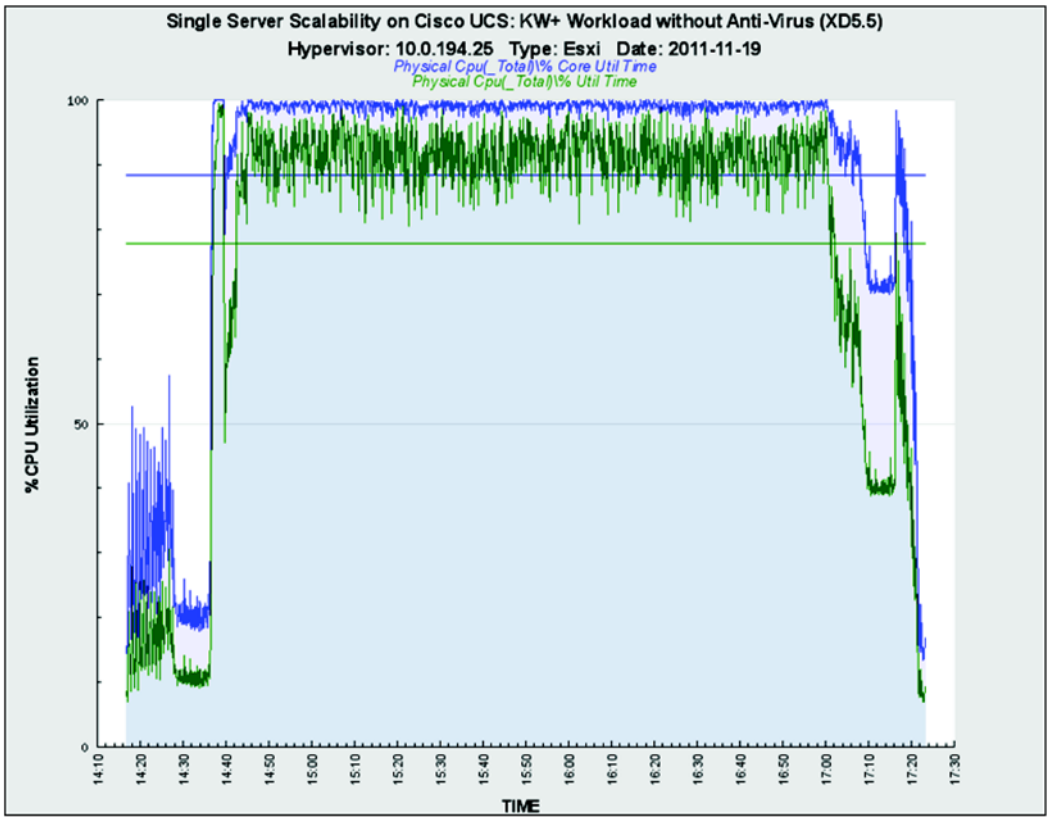
For the deployment profile detailed above, 123 VMs can be supported on a Cisco UCS B250 M2 with the following performance metrics.

Server Metrics:

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

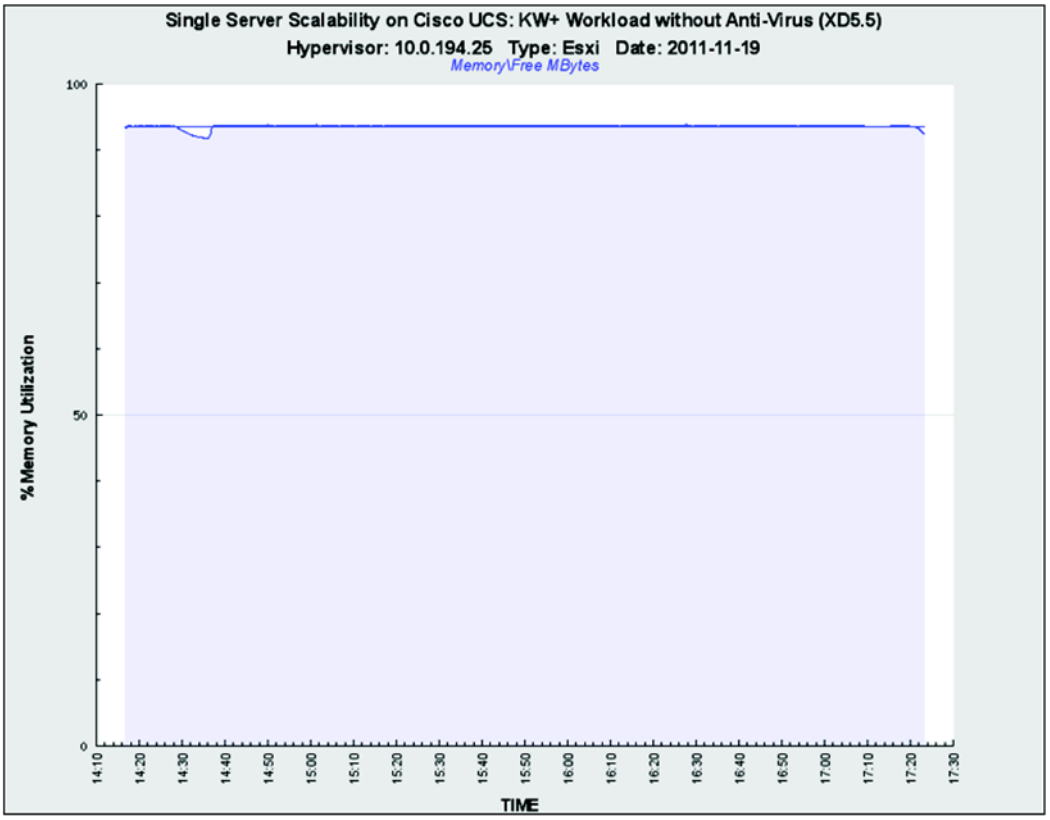
CPU Utilization

Figure 40 CPU Utilization Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)



Memory Utilization

Figure 41 Memory Utilization Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)



IO Statistics

Figure 42 IOPS Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)

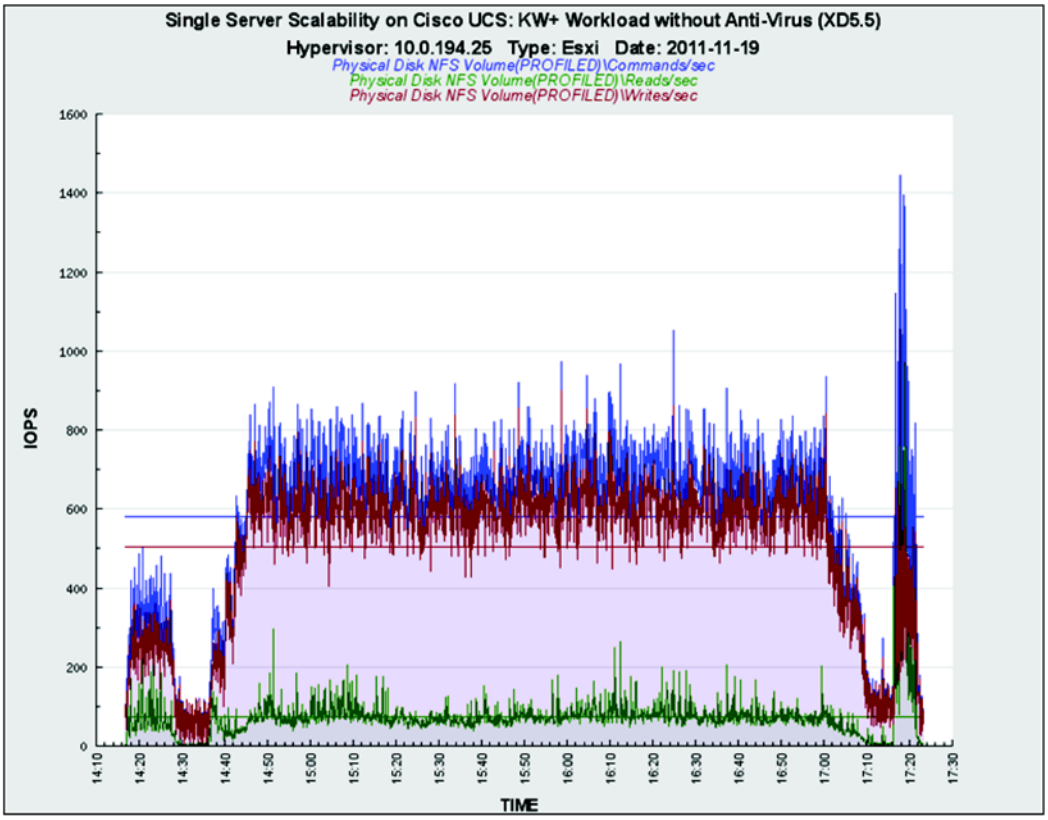


Figure 43 IO Latency Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)

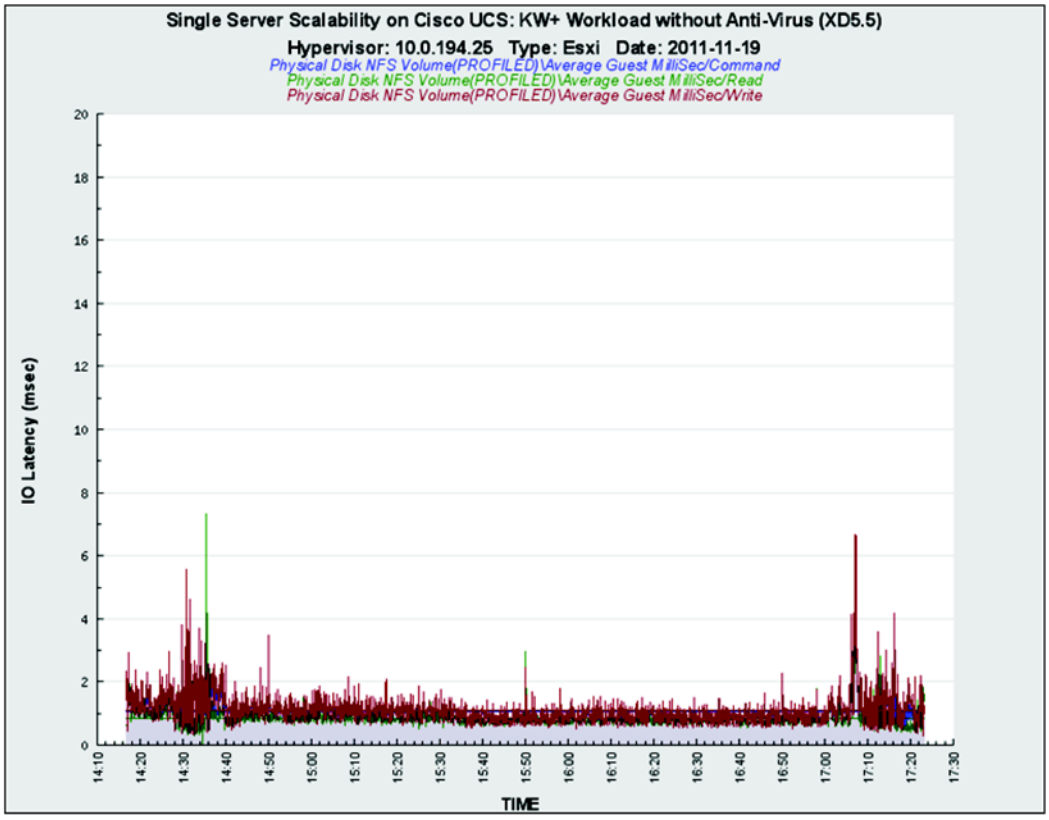
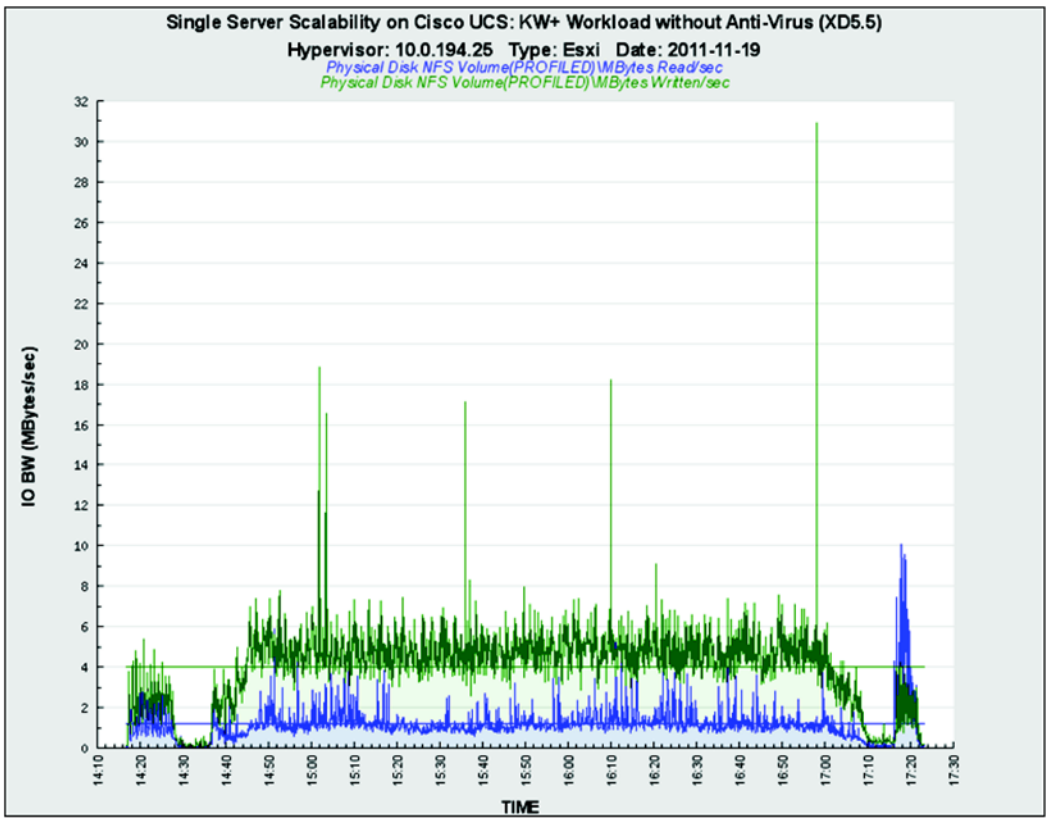
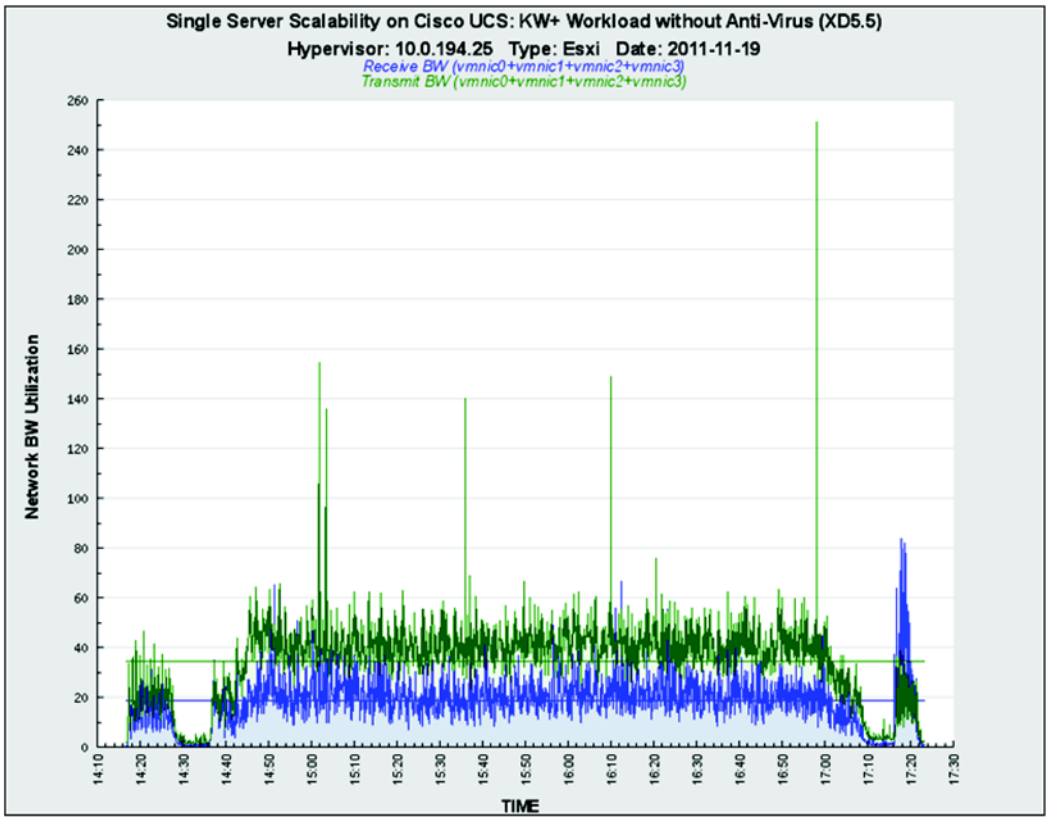


Figure 44 IO BW Utilization Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)



Network Bandwidth Usage

Figure 45 Network BW Utilization Chart for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)



Application Response Times

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

Table 17 Response Times for Baseline Profile (XD5.5/ESXi4.1/ICA/B250M2/NetApp)

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	1.2s
Outlook	5s	2.8s
Word	5s	0.8s
Excel	5s	0.9s
PowerPoint	5s	0.6s

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test
Internet Explorer	5s	0.9s
Acrobat	5s	0.7s

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 system are provided in the Table 18 below.

Table 18 WAN Capacity Planning

Desktop Virtualization	Workload Profile	HVD Profile	Storage	UCS Server
Objective: Understanding the bandwidth (BW) characteristics of a Cisco KW+ workload				
XenDesktop5 on ESXi 4.1	Cisco Knowledge Worker+	Win 7 32b (1.5G, 24G, 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				
XenDesktop5 on ESXi 4.1	Video-only	Win 7 32b (1.5G, 24G, 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				
XenDesktop5 on ESXi 4.1	Cisco Knowledge Worker+	Win 7 32b (1.5G, 24G, 1vCPU)	T1 with 80ms of latency	B50 M2 (2 x 6 core X5680 @3.33 GHz with 192G of memory)
Objective: Impact of Bandwidth Optimization using Cisco WAAS				
XenDesktop5.5 on ESXi 5	Cisco Knowledge Worker+	Win 7 32b (1.5G, 24G, 1vCPU) Dedicated Desktop	T1 with 80ms of latency	B50 M2 (2 x 6 core X5680 @3.33 GHz with 192G of memory)

Validation Methodology

The methodology used for characterizing VXI 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 placed 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 VDI workload – Cisco KW+ workload

This section focuses on the bandwidth characteristics of a typical VDI workload using Cisco KW+ as an example. The workload profile used is a critical factor for any performance related characterization of a Cisco Virtual Workspace system, including minimum bandwidth required for remotely displaying the virtual desktop events with good user experience. As the performance data can vary depending on the workload profile used, it is important to do a similar assessment in the customer environment, using a workload that is representative of their user base, not only in terms of applications but also with respect to usage patterns. However, the Cisco KW+ workload is a 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 a 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 VDI 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

- XenDesktop5 on ESXi 4.1
- HVD Profile:

  - Windows 7 32b with 1.5G memory
  - Display protocol: ICA
  - 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 B200 M2 with 96G of RAM – server was running at minimal loads during this test
- For this test, a single HVD was used from a branch site, across a T1 WAN link. Delay of 80ms was injected but no jitter

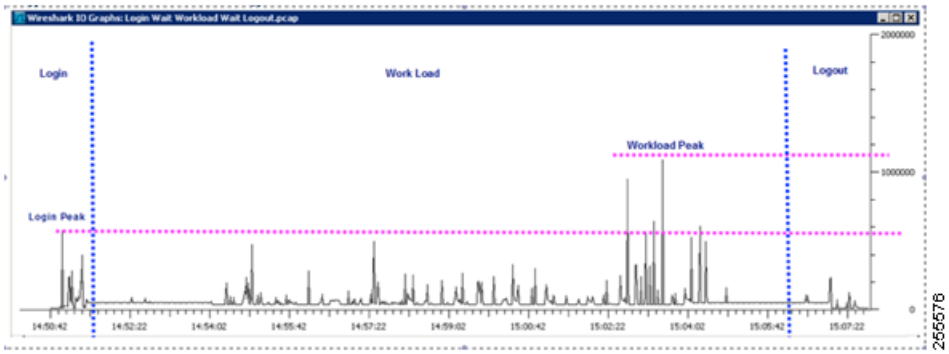
User Experience/Application Response Times

For this test, the user experience was observed over multiple iterations of a test run using an automated workload with the session experience captured on WebEx for additional review and analysis. In this particular test with a single 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. Over the course of the testing, it also became obvious that certain activities within the applications in the workload are more susceptible to limited bandwidth and therefore careful attention was paid to these areas. Examples of this include viewing a PowerPoint in Slide Show mode and composing an email in Outlook. When bandwidth restriction starts impacting UE, the information on a PowerPoint slide maybe 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 case, the bandwidth data measured should be more accurate though it is a subjective measurement.

Summary of Test Results

Bandwidth – Peak and Average

Figure 46 Bandwidth Utilization for an ICA session with Cisco KW+ workload



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 VDI session with no other traffic on the link other than minimal control traffic and the graph above is filtered view to show just the VDI session traffic. The information also shows the bandwidth utilization during

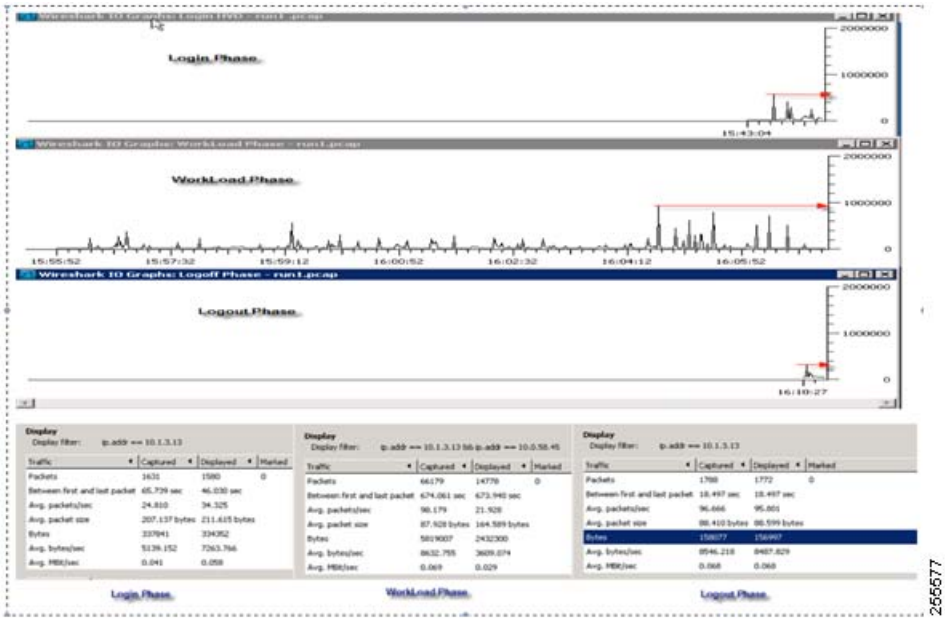
when the user first logs into a VDI session and when the user logs out. The peak bandwidth utilized in each phase is summarized in the table below. Note that the workload peaks are the highest, hitting T1 speeds, followed by the login phase. Logout phase seems to have the least BW impact among the three 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 based on the T1 bandwidth 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. It is important to note that user experience did not suffer during the workload peak though it may have been limited by the T1 link.

Table 19 Peak Bandwidth for a single VDI session using ICA 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	600 kbps	500 kbps	600 kbps	567 kbps
Workload	900 kbps	1 Mbps	1 Mbps	967kbps
Logout	300 kbps	300 kbps	280 kbps	293 kbps

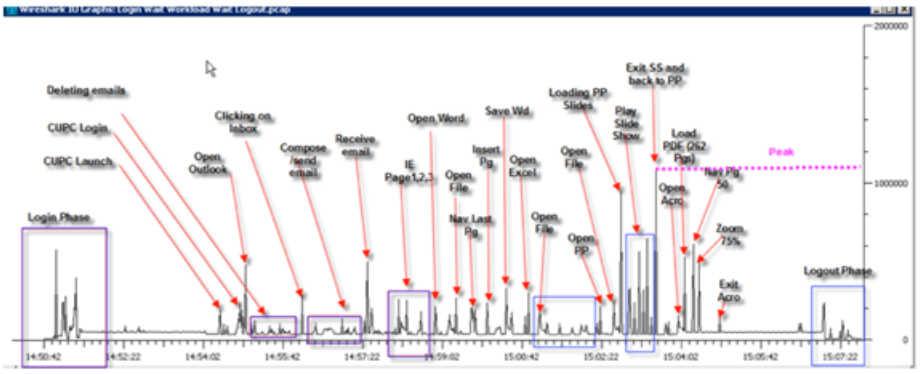
The above data for a single user using a given workload can now be used in conjunction with the minimum BW data to define the bandwidth range that provides good UE – this data is key to the sizing the WAN link for a branch Cisco Virtual Workspace deployment.

Figure 47 Bandwidth Utilization for an ICA session with Cisco KW+ workload - Alternate View



The above figure provides an alternate view on the BW utilization for each phase as well as shows the average utilization (table below graph) for the short, single iteration run that is shown in the graph above.

Figure 48 Detailed Application View of Cisco KW+ workload

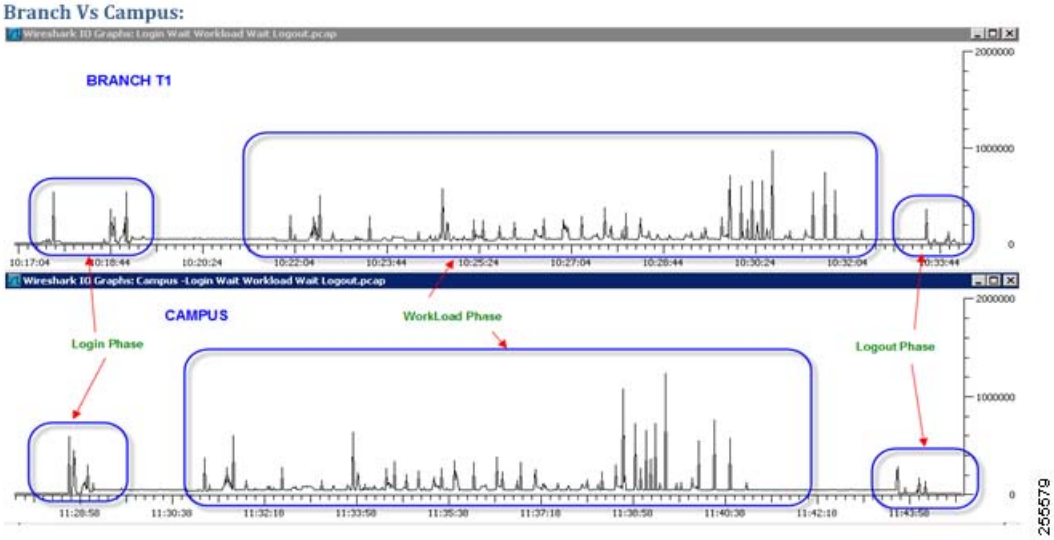


In this figure, the workload phase is further detailed in terms of the applications and activities within the workload. This shows both the absolute and relative BW impact that a given operation within the HVD has when it is remotely displayed to the user.

Note Peak bandwidth during this workload is seen from PowerPoint, followed by Acrobat. 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 bandwidth congestion.

Branch versus Campus

Figure 49 Branch vs. Campus View of Bandwidth Utilization for ICA



The above figure shows the peak bandwidth usage when going from a T1 with a single user (and delay of 80ms) to a Campus with 100Mbps+ bandwidth (and no delay). Above graph clearly shows that the peak bandwidth seen for the same workload is actually higher and if you're sizing the WAN link to

accommodate the peaks or to an X% of that peak, its important to determine the peak bandwidth in an environment with 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 to reflect the true peak BW during the workload phase as follows:

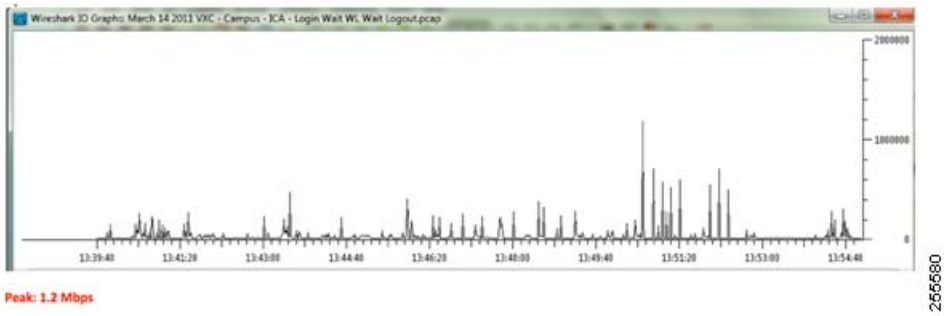
Table 20 Peak Bandwidth for a single VDI session using ICA and a Cisco KW+ workload from Campus

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

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 50 ICA Minimum BW for Cisco VXC client running Cisco KW+ workload



Using the above methodology, the minimum bandwidth for good user experience when using ICA as the display protocol is determined to be 128kbps for this workload.

Bandwidth Characteristics of a Video Only VDI workload

This section focuses on the bandwidth characteristics of a video only VDI 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 VDI session is reduced to create the congestion.

Test Environment and Setup

- XenDesktop5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory

  - Display protocol: ICA
  - Display Session Characteristics:
    - Screen Resolution: 1350 x 686
    - 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 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 was used from a branch site, across a T1 WAN link. Delay of 80ms was injected but no jitter

Summary of Test Results

Bandwidth – Peak and Average

The two figures below show the bandwidth utilization of the 1min video clip without congestion. 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.

Figure 51 Video-only Bandwidth Utilization - Unrestricted

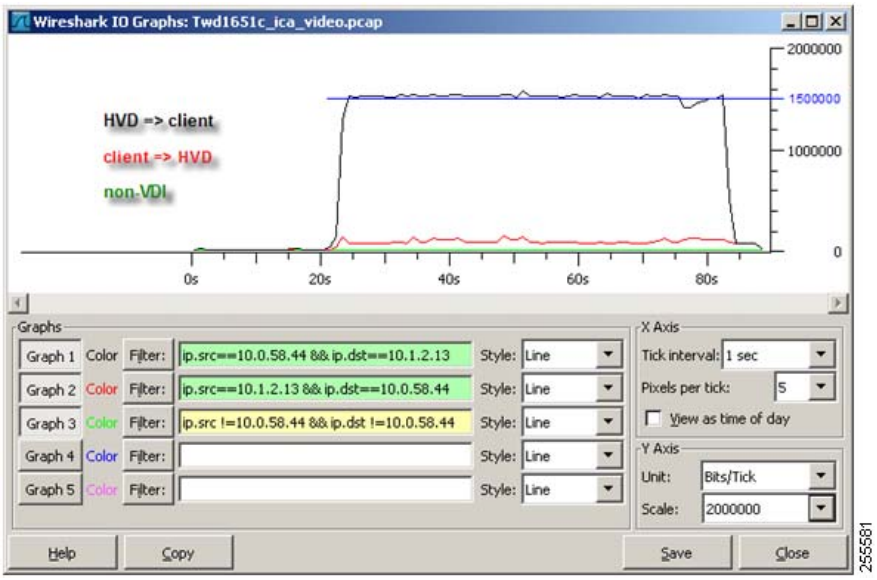
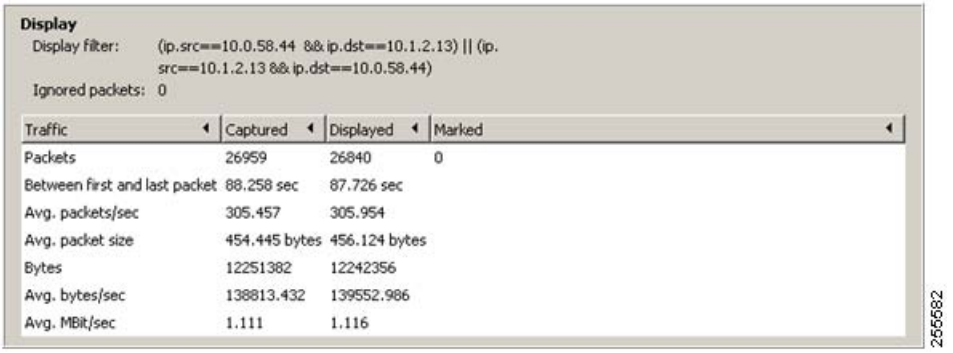


Figure 52 Wireshark Stats for Video-only Bandwidth Utilization (Unrestricted)



Display

Display filter: (ip.src==10.0.58.44 && ip.dst==10.1.2.13) || (ip.src==10.1.2.13 && ip.dst==10.0.58.44)

Ignored packets: 0

Traffic	Captured	Displayed	Marked
Packets	26959	26840	0
Between first and last packet	88.258 sec	87.726 sec	
Avg. packets/sec	305.457	305.954	
Avg. packet size	454.445 bytes	456.124 bytes	
Bytes	12251382	12242356	
Avg. bytes/sec	138813.432	139552.986	
Avg. MBit/sec	1.111	1.116	

255582

Minimum Bandwidth

The two figures below show the bandwidth utilization of the 1min video clip with congestion. 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 1024kbps. However, the user experience, both video and audio quality became unacceptable at this rate. Video was choppy, difficult to understand and audio was out-of-sync with the video. Based on this, minimum bandwidth required is a 1024kbps + 64kbps (1 64kbps timeslot) = 1088kbps approximately for a 1min, Standard Definition (640x360) video clip.

Figure 53 Video-only Bandwidth Utilization – Bandwidth Restricted

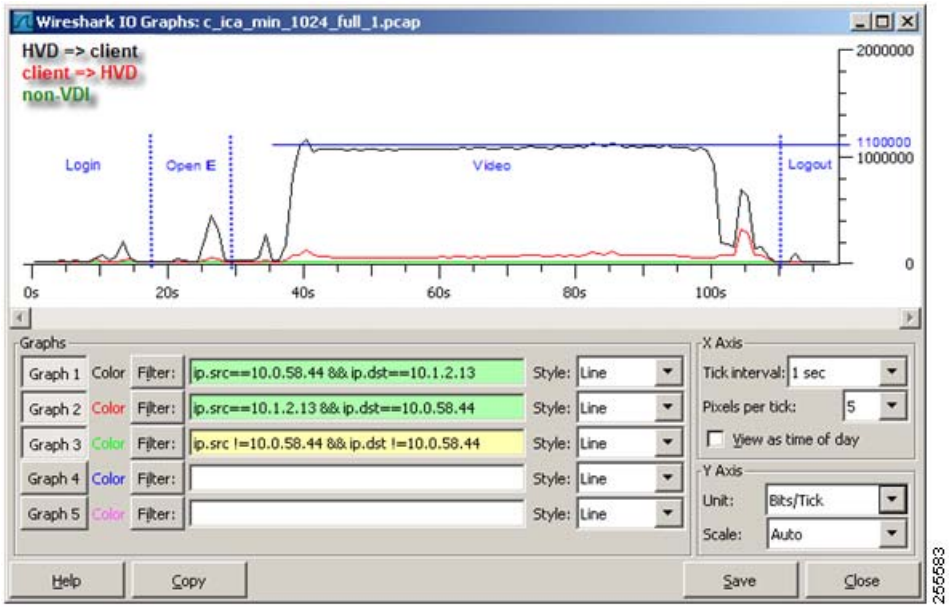
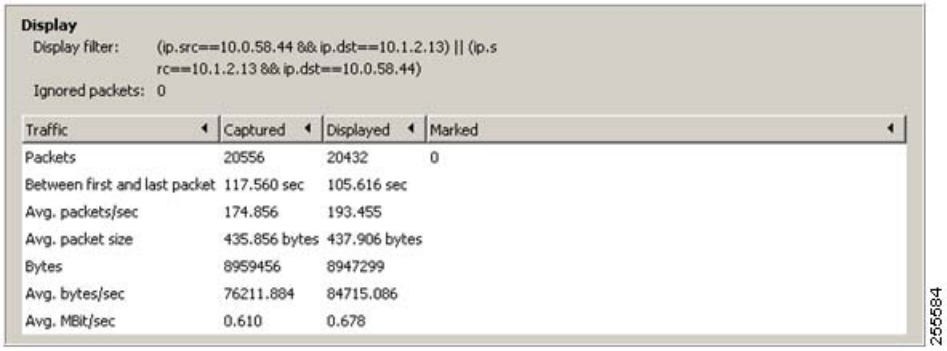


Figure 54 Wireshark Stats for Video-only Bandwidth Utilization (Restricted)



Impact of Protocol Adaptiveness on Server Performance

For large branch based VDI deployments, it is important to understand whether the adaptive nature of the display protocols has any impact on the server hosting the virtual desktops. Typically, server scale and performance benchmarking is done without any constraints to the bandwidth available across the display session. Therefore, the objective here is to determine the impact of display protocol adaptiveness on a single server when all users are in branch sites and experiencing congestion on their WAN links.

Test Environment and Setup

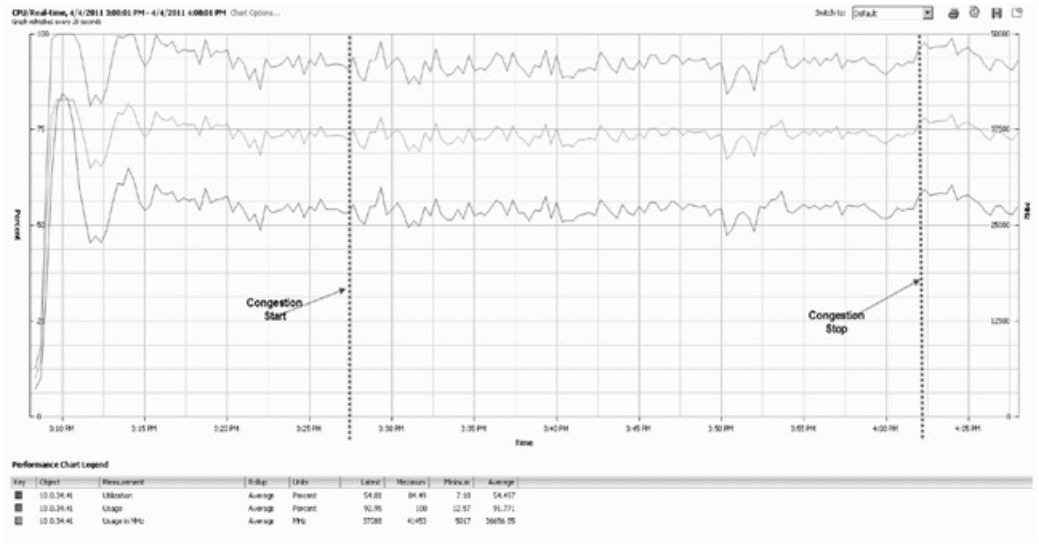
- XenDesktop5 on ESXi 4.1
- HVD Profile:
  - Windows 7 32b with 1.5G memory
  - Display protocol: ICA

  - 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

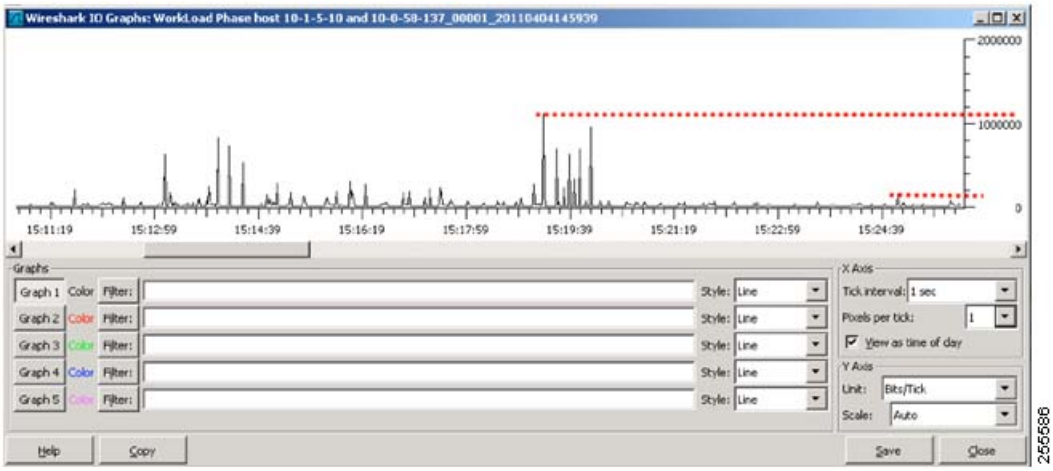
The graph below shows the CPU utilization on a Cisco UCS B250 M2 server hosting 80 HVDs where all users are in branch networks across the Cisco Virtual Workspace network. When the CPU utilization reaches a steady state of ~90 utilization, congestion is introduced on the WAN links using a traffic generator. The results show no significant impact on the server performance as the sessions adapt down to use less bandwidth.

Figure 55 Impact of Congestion on Server/Compute with ICA



The figure below shows the bandwidth utilization for one of the sessions hosted on the server above. Note that at ~ 3:19:39 pm, the session was peaking at over 1Mbps in line with results of the peak BW data presented earlier. However, once congestion as introduced, this session along with others have adapted down to well below 200kbps, again in line with the min. BW of ~128kbps determined earlier. The user experience in terms of application response times were measured for each application across all 80 sessions and were well within the acceptable range.

Figure 56 Per HVD view during congestion



Key Takeaways

Minimum bandwidth required for ICA with specified workload is 128kbps and the peak is ~1.3Mbps. 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 Acrobat are the biggest bandwidth consumers in the specified workload.

Impact of Bandwidth Optimization using Cisco WAAS

The objective of this testing is to characterize the bandwidth and performance improvements that Cisco Wide Area Application Services (WAAS) can provide for VXI branch deployments based on Citrix XenDesktop. Cisco WAAS is a comprehensive WAN optimization solution that minimizes bandwidth consumption, accelerates applications over the WAN and delivers video to the branch office while maintaining LAN-like application performance. Cisco WAAS is deployed on either end of the WAN link and can optimize Citrix ICA session traffic without disabling Citrix's native encryption or compression through a combination of technologies such as context aware data redundancy elimination (DRE), session based compression and other algorithms to reduce bandwidth consumption and improve user experience for branch users by reducing the overall effects of WAN.

In the Cisco Virtual Workspace system, using the Cisco KW+ workload, deploying WAAS results in a 26% reduction in per-session bandwidth, thereby increasing the number of virtual desktop sessions that can be supported across a given WAN link.

For this testing, 12 users were deployed at a branch site across a T1 WAN link and the bandwidth per session was measured on the WAN link with and without WAAS. Cisco KW+ defined earlier in the document was the workload running on the desktop. Latency of 80ms was present on the WAN link in both cases. No changes were made to the Citrix environment to accommodate this testing.

Results Analysis Summary

Testing done in the Cisco Virtual Workspace system shows that using Cisco WAAS can significantly reduce the Bandwidth required on the WAN link by 26%+ for a single ICA session. For a branch site with a number of virtual desktops and multiple ICA sessions, this can reduce the WAN link costs and overall costs of your virtual desktop deployment. See table below for more details.

Table 21 Bandwidth Optimization using Cisco WAAS

	Bytes Transferred Without WAAS	Bytes Transferred With WAAS	Bandwidth reduction on WAN link
Single ICA Session (CGP Enabled)	2,621,661	1,924,812	26.6%

- It is important to note that the above bandwidth savings is in addition to any ICA session level optimization that may have already reduced the bandwidth required per session. Display protocols such as ICA tend to adapt based on the available bandwidth, latency etc. and the the setup for this testing was across a T1 link with 12 users and 80ms of WAN latency. Due to this setup, the ICA session would already be operating at a lower bandwidth when WAAS attempts to further optimize it would have already compressed this stream and WAAS will further optimize this stream. As such, bandwidth saving with WAAS represents savings beyond ICA session level optimization.
- The 26% bandwidth reduction shown in the table above is strictly for ICA session traffic. However, virtual desktop users could have print traffic and streaming video traffic that may need to traverse the WAN link. With these additional traffic types, bandwidth reduction provided by WAAS can be higher than 26%. Cisco KW+ workload used in this test did include Internet Explorer browsing to sites with flash images but it did not include streaming video or print traffic.
- Testing also showed that the user experience improves when using WAAS. This was reflected in the response times associated with application launch as well as the time taken to launch a session from the branch site. The response times with WAAS improved by an average of 18.5%.

Design & Deployment Considerations

- A key design consideration when deploying Cisco WAAS is that it can provide bandwidth savings for different types of traffic in addition to the virtual desktop session traffic. So if you branches have a mixed user base of virtual desktop users and traditional desktop users, there is added benefit in deploying WAAS since it could optimize both types of traffic.
- For any deployment with a number of branch sites, a single headend WAAS at the hub campus site can be used for all branch sites. This headend WAAS is highly scalable and can lower the overall TCO as the number of branch sites grow. WAAS can also be virtualized and deployed as a VM in the Enterprise data center, further extending the TCO benefits of consolidation in the data center to include not only servers and virtual desktops but also network services elements such as WAAS.
- By reducing the bandwidth required per session, WAAS can increase the number of virtual desktops sessions that a given WAN link can support with good or better user experience. Without WAAS, testing in the Cisco Virtual Workspace system using a Cisco KW+ workload showed that a single ICA session requires a minimum bandwidth of 128 kbps on average to maintain good user experience. Based on this number, a T1 link can support 12 virtual desktops concurrently – however with WAAS reducing the ICA bandwidth requirements per session by 26%, it should now be

possible to support 16 concurrent users, representing a 33% increase in the number of virtual desktop sessions supported across that same link. Densities higher than 33% is also possible depending on the workload used but here it based on using Cisco KW+ workload.

- WAAS can optimize across sessions by caching data. A video stream destined to a user can be cached by WAAS and streamed to multiple users.
- As stated above, WAAS optimization results were obtained in the Cisco Virtual Workspace system using a specific workload profile. WAAS optimization results will vary depending on the workload and specific WAN environment, so it is recommended that validation and due diligence is performed in the actual deployment environment when performing network capacity planning and characterizing the optimization attainable with WAAS

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.

Test Profile

Desktop Virtualization

- XenDesktop 5.5 using MCS
- Display Protocol – ICA
  - Multistream ICA disabled (default)
  - CGP for session reliability enabled (default)
- Pooled Desktops – Static

Hypervisor

VMware ESXi 5.0

Virtual Desktop Configuration

- Windows 7 32b
- 1.5G of RAM allocated per desktop
- 24G disk configured per desktop
- 1 vCPU per desktop
- Dedicated desktop
- Display Session Characteristics
  - Screen Resolution: 1366 x 768 (Large Window)
  - Color Depth: 16bit
  - Windows Optimized for Best Performance

WAAS Configuration

- WAAS hardware for branch and headend: WAVE-674 running 4.5.1
- WAAS Remote-Desktop policy set to TCP Flow Optimizations (TFO) with Data Redundancy Elimination (DRE), Bidirectional Cache, LZ and Citrix ICA Application Optimization (AO)
- Virtual Central Manager version 4.5.1 was used to centrally manage WAAS components

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

- Microsoft Office 2007 Applications
- Internet Explorer
- Adobe Acrobat
- Cisco Unified Personal Communicator 8.5.1 in desk phone mode
- Applications were loaded in sequence on each desktop

Data Collection/Test Tool

- Workload Generation Tool from Scapa Test Technologies
- Response times measured using Scapa TPP as outlined in an earlier section
- Data is captured using Acterna WAN analyzer and graphed for Login, Workload & Logout phases using wireshark
- Subjective user experience is also monitored across one user session

In the next section, we look at the graphs showing the bandwidth usage per session with and without WAAS.

Bandwidth Utilization Charts

As stated before, there is a 26%+ reduction in bandwidth consumption by using WAAS. The charts below show the bandwidth profile for one user (same user in both cases) among the 12 users on the WAN link. The second chart clearly shows that the peaks in bandwidth usage are suppressed with WAAS.

Figure 57 Bandwidth Charts without WAAS for 1 user (XD5.5/ESXi5/ICA)

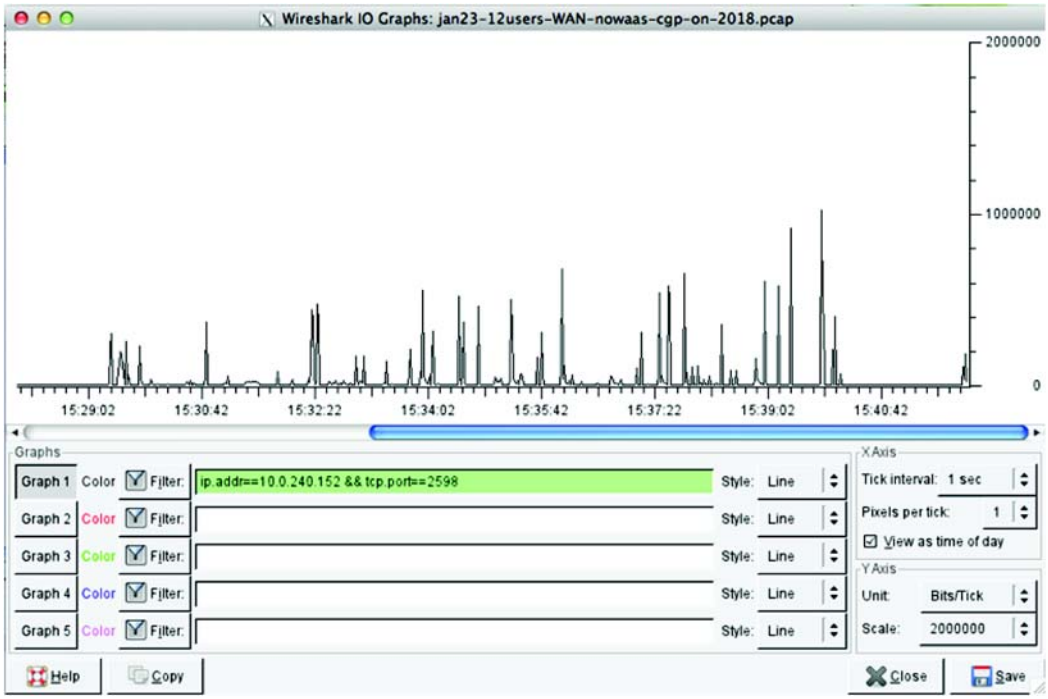
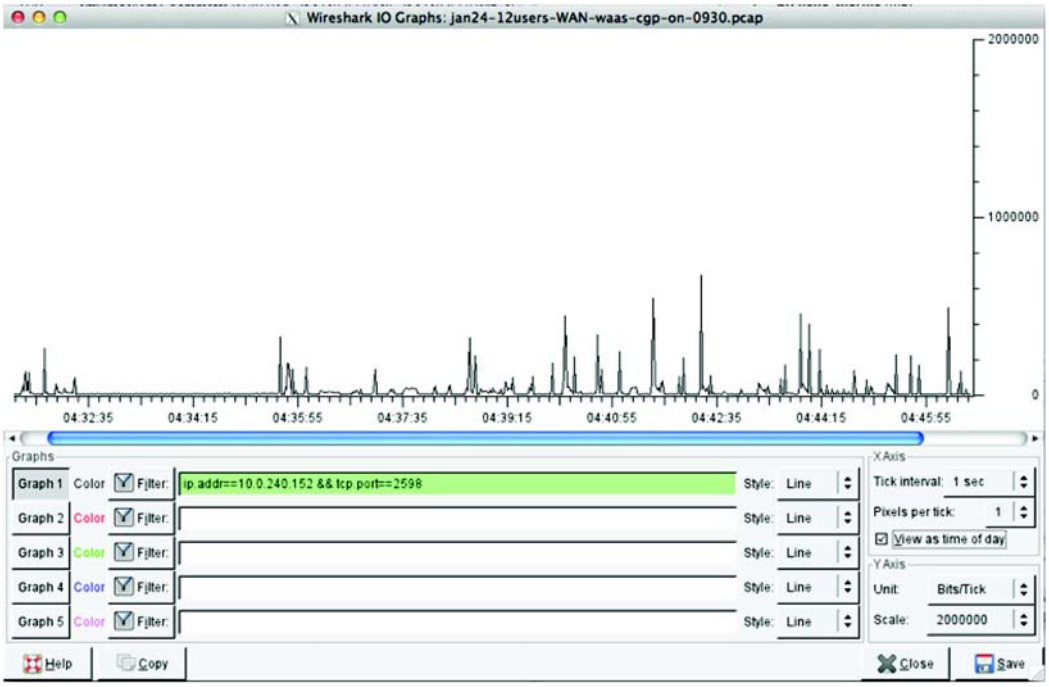


Figure 58 Bandwidth Charts with WAAS for 1 user (XD5.5/ESXi5/ICA)



Application Response Times

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

Table 22 Response Times with and without WAAS (XD5.5/ESXi5/ICA)

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured during Test (With WAAS/Without WAAS)	Response Time Improvement with WAAS
Cisco Unified Personal Communicator 8.5 in deskphone mode	5s	898ms/1828ms	.93s = 51%
Outlook	5s	1698ms/1851ms	.153s = 8.3%
Word	5s	681ms/715ms	.055s = 7.7%
Excel	5s	748ms/803ms	.055s = 6.9%
PowerPoint	5s	440ms/527ms	.087s = 16.5%
Internet Explorer	5s	728ms/749ms	.021s = 2.8%
Acrobat	5s	444ms/695ms	.251s= 36%

Note Average Response Time Improvement = 18.5%

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 23 below.

Table 23 Summary of Applications

Objective	Server Model	Storage	Desktop Virtualization Profile	HVD Profile
Scale and Performance characterization of Cisco Jabber for Windows with Citrix XenDesktop and XenApp	Cisco UCS B200 M3 Blade Server with 384 GB of memory	NFS on NetApp FAS 3170	• Citrix XenDesktop 5.6FP1 (MCS) on VMware ESXi 5.1 • Citrix XenApp 6.5 on VMware 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 256G of memory	NFS on NetApp FAS 3170	N/A - See test profile for more detail.	Microsoft Windows 7 32b with 2G 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 on Citrix XenDesktop and XenApp

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 Citrix XenDesktop (XD) and XenApp (XA).

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 (physical and virtual), 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) and hosted shared desktops (HSD) based on Citrix XenDesktop and XenApp respectively. 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: Cisco Virtualization Experience Infrastructure Smart Solution 2.6 with Citrix XenDesktop 5.6.

Note With Citrix XenApp, Cisco Jabber for Windows is supported for Citrix XenApp session virtualization or XenApp Hosted Shared Desktops (also known as published desktops or shared hosted desktops). Support for XenApp Application virtualization is currently not supported.

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.

When using Citrix XenDesktop, 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 150 HVD and 150 HSD users using Cisco Jabber. Hosted Virtual desktops were deployed on a Cisco UCS B200 M3 server using Citrix XenDesktop 5.6FP1 Machine Creation Services (MCS) and a separate Cisco UCS B200 M3 was used for the Citrix XenApp servers hosting the 150 HSD users. The usage profile used for the Citrix XenDesktop and XenApp testing is defined in the Workload Profile section of Table 3 and Table 8 respectively. For both Citrix XenDesktop and Citrix XenApp, testing involved 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. Performance data using Cisco Jabber for IM and presence are presented later in this document and similar data with Cisco Jabber used for telephony should be available in a future release of the 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 two sections, the results from the Cisco Jabber Application characterization testing done in a Citrix XenDesktop and XenApp HSD environment are provided.

Validation Overview and Results – Citrix XenDesktop

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 Citrix XenDesktop based 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 32-bit, 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 Citrix XD desktops. As each user logs into their Citrix virtual 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 charts based on 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 24 Resource Utilization on a Cisco UCS B200 M3 server with 150 Citrix virtual desktops running Cisco Jabber

	Launch & Login	Steady State	Desktop Session Logout
CPU Utilization-Avg.	29.09	21.73	17.69
CPU Utilization-Peak	43.52	29.39	42.12
Memory Allocated (%)	-	77.89	-
Read-Avg	24.19	12.62	24.01

	Launch & Login	Steady State	Desktop Session Logout
Read-Peak	101.03	779.54	54.10
Write-Avg	321.86	295.71	282.36
Write-Peak	678.52	728.35	707.77
Read-Latency-Avg.	1.76	1.19	1.76
Read-Latency-Peak	3.07	2.86	3.07
Write-Latency-Avg.	1.40	1.37	1.40
Write-Latency-Peak	3.90	5.94	3.90
Network BW (Mbps)-Avg.	38.91	19.24	29.64
Network BW (Mbps)-Peak	82.20	65.77	148.18

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 Configuration and Setup 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 25 read IOPS and 325 write IOPS for a combined total of 350 average IOPS. Peak I/O load generated is approximately 100 peak read IOPS and 725 peak write IOPs, for a total of ~825 peak IOPS. Excluding the peak read I/O data from Steady State as it is momentary (see charts) 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 Reads) 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 1/5 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.

The network bandwidth utilization includes all traffic sent and received by 150 desktops running on the server and includes NFS storage traffic. Since this is a Cisco Jabber-only workload, a majority of the network traffic from the server is also Cisco Jabber related. For a breakdown of the storage NFS traffic

vs. the total network traffic, see Performance Charts below. The average network bandwidth utilization is 20Mbps during steady state desktop use and ~40Mbps during the 10min+ window when users are launching and logging into Cisco Jabber across all 150 desktops.

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 25 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 IOPSPeak = ~1/5 for Read/Write IOPS	Derived using the following calculation: • Average = ~25R/325W IOPS/150 users= ~1R/2W IOPS/user • Peak =~100R/725W IOPS/150 users= ~1R/5W IOPS/user
Network BW	Average Network BW utilization = ~200kbps	Derived using the following calculation: Average = ~30 Mbps /150 users = ~200kbps/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.

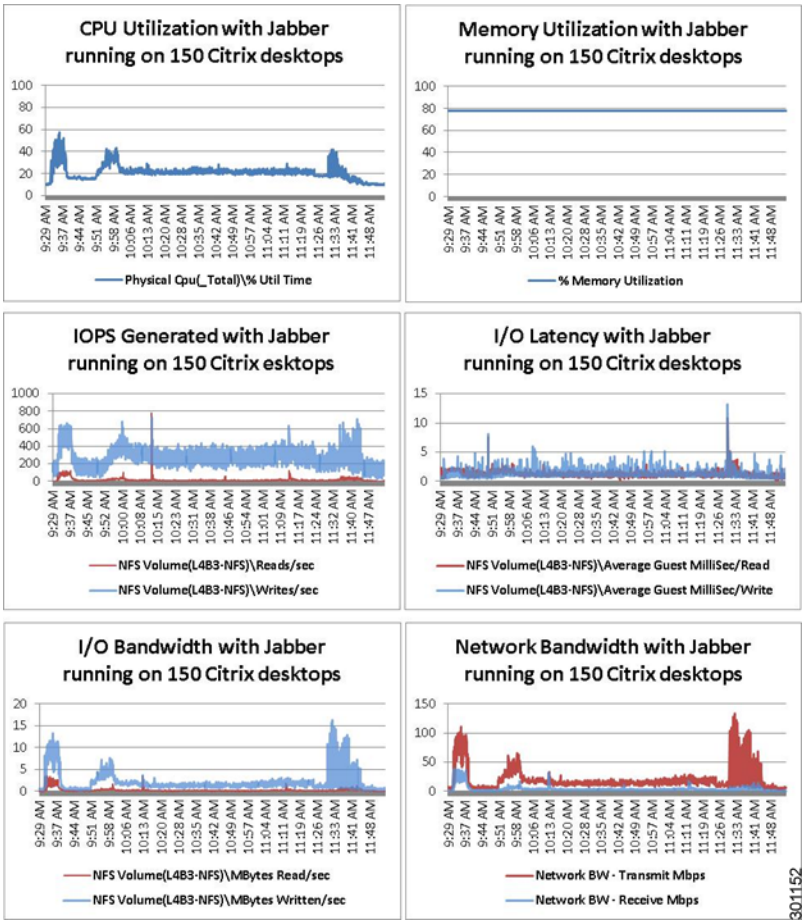
Table 26 Configuration and Setup used in Cisco Jabber testing across 150 virtual desktops

Desktop Virtualization	Citrix XenDesktop 5.6FP1
UCS Server	UCS B200 M3 with Dual Eight Core Intel®Xeon® CPU E5-2690@ 2.9GHz with 384GB of memory
Hypervisor	VMware ESXi 5.1
Storage	NetApp FAS 3170
Virtual Desktop Configuration	Windows 7 32-bit desktops with 2G of RAM and 20G disk, 1 vCPU, No memory reservation
Cisco Jabber for Windows	9.1.3

Desktop Virtualization	Citrix XenDesktop 5.6FP1
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 Citrix 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 59 Performance Charts for a Cisco UCS B200 M3 with 150 Citrix desktops using 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 27 Response Times for 150 XD users on Cisco UCS B200M3 with XD5.6.2/ESX5.1/ICA/NetApp

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 150 Citrix XD users on UCS B200M3
Cisco Jabber for Windows	5s	0.6s

Validation Overview and Results – Citrix XenApp

The goal of this testing is to characterize the scale and performance of Cisco Jabber application being used by 150 Windows users whose desktops sessions were hosted in the data center. For the testing, eight XenApp server VMs were deployed on a Cisco UCS B200 M3 server with 384GB of memory. Cisco Jabber for Windows client was installed on the XenApp Server VMs running Windows 2008 R2 SP1 server OS with each VM allocated 4vCPUs and 16GB of RAM.

Test was started by using a Test Tool representing the end users to initiate and login into 150 Citrix XA session based desktops. As each user logs into their session based desktop, each user launches Cisco Jabber, logs in and executes the Cisco Jabber workload as described in the Test Configuration and Setup section below. The Cisco Jabber-only workload is kept running for a minimum of 2 hours and represents 150 users in steady state use of Cisco Jabber. At the end of this time period, users log out of their session based desktops which also results in users logging off and quitting Cisco Jabber. 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 charts based on the data collected from the Cisco UCS server are provided in the Performance Charts section below. The charts shows the Cisco Jabber resource utilization of 150 session based 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 28 Resource Utilization on a Cisco UCS B200 M3 server with 150 Citrix XA session based desktops running Cisco Jabber

	Launch & Login	Steady State	Desktop Session Logout
CPU Utilization-Avg.	21.99	7.38	7.35
CPU Utilization-Peak	40.99	13.85	30.76
Memory Allocated (%)	-	33.98	-
Read-Avg	10.16	0.06	1.02
Read-Peak	44.27	7.28	15.62
Write-Avg	140.85	78.53	115.81
Write-Peak	277.45	179.28	468.86
Read-Latency-Avg.	3.28	1.37	1.58
Read-Latency-Peak	8.83	68.74	22.85
Write-Latency-Avg.	1.25	0.76	1.92
Write-Latency-Peak	2.73	2.41	22.85
Network BW (Mbps)-Avg.	36.35	7.89	27.92
Network BW (Mbps)-Peak	109.94	23.09	159.60

The data shows that Cisco Jabber with Citrix XenApp uses less than 10% 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 Configuration and Setup section below. During the launch and login stage, CPU utilization on the server is at ~20% (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 35% of the available memory on the server was allocated to support 150 session based desktops, distributed across 8 XA server VMs. The UCS server used in the test was deployed with 384GB of memory with 16GB of memory allocated to each XA server VM. The utilization of 35% represents the memory allocated to the 8 XA servers, 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 for each XA server and at the UCS server level to ensure that there is memory available for supporting the desktop users running on that server.

From a storage perspective, the average I/O load generated by 150 Cisco Jabber users for the given workload profile is approximately 10 read IOPS and 150 write IOPS for a combined total of 160 average IOPS. Peak I/O load generated is approximately 45 peak read IOPS and 500 peak write IOPs, for a total of ~545 peak IOPS.

The I/O activity in the Logout stage involves logging off from Cisco Jabber server, closing Cisco Jabber application and logging off from their session based 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 Reads) 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 1/1 for average read/write IOPS and 1/4 for peak read/write IOPS.

I/O latency experienced by the Microsoft Guest OS running on the XA server VMs is well below the acceptable threshold of 20ms (average) throughout the test. An intermitted peak of 68msec was seen but happens only once during the 2 hour run so considering this as an anomaly, particularly with average being well below 20msec.

The network bandwidth utilization includes all traffic sent and received by the 150 XA session based desktops running on the server and includes NFS storage traffic. Since this is a Cisco Jabber-only workload, a majority of the network traffic from the server is also Cisco Jabber related. For a breakdown of the storage NFS traffic vs. the total network traffic, see Performance Charts below. The average network bandwidth utilization is less than 10Mbps during steady state desktop use and ~40Mbps during the 10min+ window when users are launching and logging into Cisco Jabber across all 150 desktops.

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 29 Compute, Storage and Performance Requirements for a single Citrix XA desktop running Cisco Jabber

Compute	Average = ~31 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) = 10% = .10x 46.4GHz = 4.6GHz Average CPU cycles needed per desktop = 4.6GHz/150 = 31 MHz
Memory	2GB per user	Assuming no memory over-subscription

Storage I/O	Average = ~1/1 for Read/Write IOPS Peak = ~1/4 for Read/Write IOPS	Derived using the following calculation: Average = ~10R/150W IOPS/150 users = ~1R/1W IOPS/user Peak =~45R/500W IOPS/150 users = ~1R/4W IOPS/user
Network BW	Average Network BW utilization = ~160kbps	Derived using the following calculation: Average = ~24 Mbps /150 users = ~160kbps/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 session based 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.

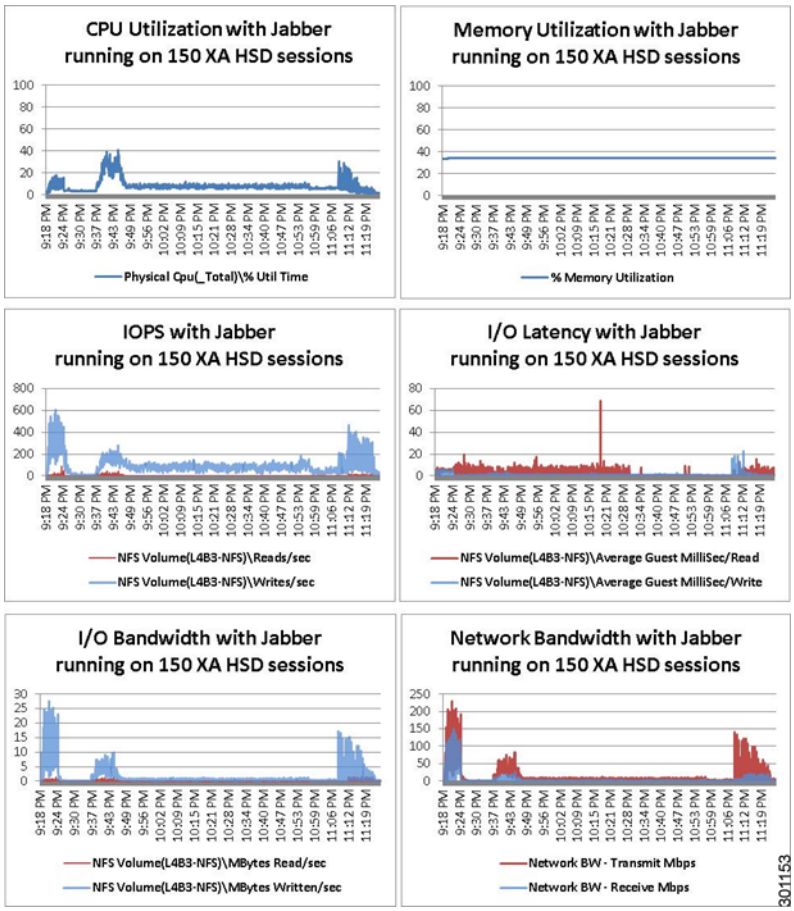
Table 30 Configuration and Setup used in Cisco Jabber testing across 150 Citrix XA desktops

Desktop Virtualization	Citrix XenApp 6.5
UCS Server	UCS B200 M3 with Dual Eight Core Intel® Xeon® CPU E5-2690@ 2.9GHz with 384GB of memory
Hypervisor	VMware ESXi 5.1
Storage	NetApp FAS 3170
XenApp Server Virtual Machine Configuration	Eight Windows 2008 R2 SP1 Server VM with 4vCPUs, 16G of RAM and 80G disk, no memory reservation
Cisco Jabber for Windows	9.1.3

Desktop Virtualization	Citrix XenApp 6.5
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 workload runs on all 150 session based desktops • The exact steps performed during testing are outlined below: – Launch and Login to 150 Citrix session based desktops – Wait until Desktop Login phase completes – Start the workload using test tool; tool will randomize the start of the workload so that the workload is randomized across the 150 desktop sessions – 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 desktop session 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 60 Performance Charts for a Cisco UCS B200 M3 with 150 Citrix XA desktops using 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 31 Response Times for 150 XA users on Cisco UCS B200M3 with XA6.5/ESX5.1/ICA/NetApp

Applications	Maximum Acceptable Startup Times (Success Criteria)	Average Startup Times Measured for 150 Citrix XA users on UCS B200M3
Cisco Jabber for Windows	5s	0.6s

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 Citrix XenDesktop or XenApp desktops.

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

A fundamental aspect of deploying Cisco Virtual Workspace Solution 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 32 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 33 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 34 provides configuration, environment and setup details used in this testing.

Table 34 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 61 Performance Charts for Cisco UCS B230 M2 with 120 Contact Center Agent desktops running Cisco CTI OS client