Scalability and economics of XenApp on Amazon cloud

Contents

Introduction

Chief Information Officers (CIOs) are looking to the cloud as the new frontier for IT services, based on its ability to deliver cost-effective, enterprise-class virtual desktop and application solutions as provided by Citrix XenDesktop®. The attractiveness of a XenDesktop, cloud-based desktop and application offering is driven by the ever-evolving demands of workforce growth, flexibility and geographical disbursement. The ability to reallocate upfront costs of a large physical hardware infrastructure purchase into an operating expense model distributed over an extended timeframe with on-demand instant access to resources is an extremely attractive proposition. The instant availability of cloud-based resources dramatically reduces the implementation time previously allocated for procuring, racking and cabling physical servers. The scalability and flexibility of addressing a dynamic shift in workforce, albeit through a merger, acquisition or simple seasonal demand, provides even more business justification for a cloud-oriented deployment. Most importantly, by delivering virtual desktops and applications through XenDesktop utilizing Amazon Web Services, CIOs are able to deliver a powerful virtualization solution with a price point as low as one penny an hour per user.

Amazon Web Services

Since early 2006, Amazon Web Services (AWS) has provided companies of all sizes with an infrastructure web services platform in the cloud. With AWS you can requisition compute power, storage, and other services–gaining access to a suite of elastic IT infrastructure services as your business demands them. You pay only for what you use, with no up-front expenses or long-term commitments, making AWS the most cost-effective way to deliver your application to your customers and clients. With AWS, you can take advantage of Amazon.com's global computing infrastructure, which is the backbone of Amazon.com's multi-billion-dollar retail business and transactional enterprise who's scalable, reliable, and secure distributed computing infrastructure has been honed for over a decade.

AWS is cost effective, dependable, flexible and comprehensive. With AWS, you pay only for what you use, with no up-front expenses or long-term commitments. The Amazon cloud is scalable, with massive compute capacity and storage. It is reliable, redundant and secure.

Amazon Virtual Private Cloud (VPC) lets you provision a private, isolated section of the AWS cloud where you can launch AWS resources in a virtual network that you define. With Amazon VPC, you can define a virtual network topology that closely resembles a traditional network that you might operate in your own datacenter. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

You can easily customize the network configuration for your Amazon VPC. For example, you can create a public-facing subnet for your webservers that has access to the Internet and place your backend systems, such as databases or application servers, in a private-facing subnet with no Internet access. You can leverage multiple layers of security, including security groups and network-access-control lists, to help manage access to Amazon EC2 instances in each subnet.

Citrix XenApp®

Citrix XenApp® 6.5 provides advanced management and scalability, a rich multimedia experience over any network, and self-service applications with universal device support ranging from laptops to smartphones. Full support for Windows Server® 2008 R2 and seamless integration with Microsoft® App-V, XenApp 6.5 provides session and application virtualization technologies that make it easy for customers to centrally manage applications using any combination of local and hosted delivery to best fit their unique requirements.

XenApp 6.5 introduces significant enhancements that simplify application management and bring unprecedented levels of scalability to increase cost savings and datacenter efficiency. XenApp gives you the ability to centrally manage heterogeneous applications and deliver Software as a Service (SaaS) to any workforce.

Delivering these desktops and applications from XenApp presents unique challenges to administrators who are accustomed to delivering traditional cloud-hosted applications or services. Being able to determine how many users can be hosted on a single Amazon Elastic Compute Cloud (Amazon EC2) instance is critical in order to achieve maximum operation efficiency.

This whitepaper examines the scalability of XenApp 6.5 on various EC2 instances. It will guide you through the selection and sizing of the different Amazon EC2 instance types, as well as the economic impact of using these instances, allowing you to further reduce cost while securely delivering applications to anyone, anywhere, anytime, on any device.

Results summary

In order to determine user scalability of XenApp servers, the Login VSI performance and scalability testing tool from Login Consultant was used to simulate real-world user interactions within Microsoft Office Professional 2010 applications. Login VSI allows for complete test automation including session launch, user action simulation and the monitoring of session performance characteristics.

The bar graph below shows the maximum number of simulated XenApp user sessions that each EC2 instance configuration can support before encountering a user-experience bottleneck. Please see the detailed results section for more information on the bottlenecks.

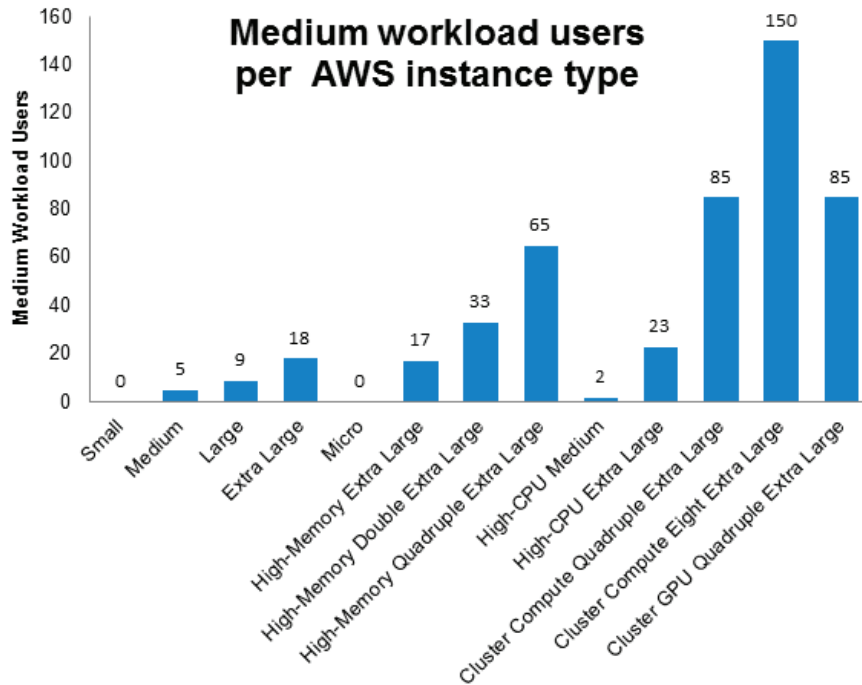Figure 1 XenApp Hosted Shared Desktops per AWS Instance Type

Table 1 AWS Compute costing per hour

Instance type	Compute units	RAM (GB)	vCPUs	East Coast Cost per hr	User Sessions	Cost per hr per user
Standard small	1	1.7	1	$0.115	0	N/A
Standard medium	2	3.7	2	$0.230	5	$0.0460
Standard large	4	7.5	2	$0.460	9	$0.0511
Standard extra large	8	15	4	$0.920	18	$0.0511
Micro 32-bit or 64-bit	1	0.613	1	$0.030	0	N/A
High-memory extra large	6.5	17.1	2	$0.570	17	$0.0335
High-memory double extra large	13	34.2	4	$1.140	33	$0.0345
High-memory quadruple extra large	26	68.4	8	$2.280	65	$0.0351
High-CPU medium	5	1.7	2	$0.285	2	$0.1425
High-CPU extra large	20	7	8	$1.140	23	$0.0495
Cluster compute quadruple extra large	33.5	23	16	$1.610	85	$0.0189
Cluster compute eight extra large	88	60.5	32	$2.970	150	$0.0198
Cluster GPU quadruple extra large	33.5	22	16	$2.600	85	$0.0306

For reference, two types of XenApp AWS sample environments are provided for an estimate of the costs of running a XenApp 6.5 farm on AWS.

These sample reference designs are for a 1,000 CCU and 10,000 CCU XenApp Farm, respectively.

Both designs use a core infrastructure that is powered on 24 hours a day, and a dynamic, elastic part that is active for 14 hours a day, covering normal eight hour work days, as well as providing ample margin for work shifting.

The design makes use of a three year term of AWS Reserved Instances instead of on-demand instances. With Amazon EC2 Reserved Instances, you pay a low, one-time fee and in turn receive a significant discount on the hourly charge for that instance. See http://aws.amazon.com/ec2/reserved-instances/ on the Amazon Web Services site for more details.

The XenApp server instances have been configured with Amazon Elastic Block Store (EBS) storage for their boot device and with local-instance storage to hold volatile data such as pagefiles and user profiles. EBS storage provides block-level storage volumes for use with Amazon EC2 instances. Amazon EBS volumes are off-instance storage that persist independently from the life of an instance. Amazon EBS provides high availability, highly reliable storage volumes, placed and automatically replicated in a specific Amazon Availability Zone.

Availability Zones are distinct locations that are engineered to be insulated from failures in other Availability Zones and provide inexpensive, low latency network connectivity to other Availability Zones in the same Region. By launching instances in separate Availability Zones, you can protect your applications from failure of a single location. Regions consist of one or more Availability Zones, are geographically dispersed, and will be in separate geographic areas or countries.

With Amazon EBS, volume storage is charged by the amount of storage allocated, and is priced at a rate of $0.10 per allocated GB per month. Amazon EBS, however, also charges $0.10 per 1 million I/O requests to a volume.

Given that each XenApp session generates on average 7-8 IOPS per user, it is advised that you use local- instance storage instead of EBS for any volatile data. Local-instance storage is an inexpensive way to provide high-speed storage and only persists during the life of the instance.

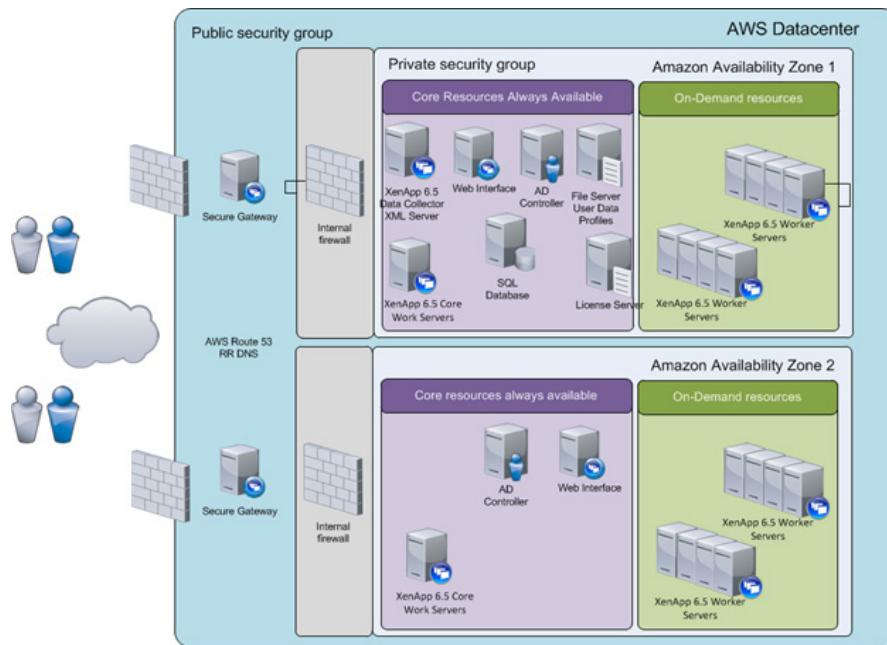Figure 2 XenApp Farm on AWS for 1,000 hosted shared desktops

Table 1 XenApp on AWS 1,000 desktops costing information
Architecture up to 1,000 users

		Costing Totals	Daily	Monthly	Annual
Cost per user per year	$71.33	XenApp Servers Elastic Capacity	$77.98	$2,371.89	$28,462.70
Cost per user per month	$5.95	Infrastructure Services @24 hrs	$40.46	$1,230.22	$14,769.36
Cost per user per day	$0.20	EBS Storage cost	$47.32	$1,439.40	$17,272.80
Cost per user per hour	$0.01	Network Data transfer cost	$4.20	$84.00	$1,008.00
Supported Logons per minute	240	Reserve Charges	$26.89	$818.06	$9,816.67
Minutes to logon all users	5	Total Recurring Cost	$213.92	$6,462.90	$71,329.53

XenApp Servers		Instance Type	Cost per hr	Extension	Total Reserve Charge
Number of users	1000				
Users per server	150				
Uptime per day (hours)	14				
Number of XA Servers	5	cc2.8xlarge	$1.114	5.57	$13,550
				3 Year Term - Reserved Instances	

Infrastructure Services		Instance Type		Cost per hr	Extension	24hrs	Total Reserve
Number of core XA servers	2	m2.xlarge	$0.140	0.28	$1,550		
Data Collector	1	m2.xlarge		$0.140	0.14	$1,550	Heavy	$3,100
SQL Server	1	m2.xlarge		$0.140	0.14	$1,550	Heavy	$1,550
Web Interface	2	m1.xlarge		$0.132	0.264	$1,200	Heavy	$1,550
License Servers	1	m1.medium		$0.066	0.066	$600	Heavy	$2,400
Active Directory Servers	2	m1.large		$0.132	0.264	$1,200	Heavy	$600
Windows File Server	1	c1.xlarge		$0.400	0.4	$3,100	Heavy	$2,400
Secure Gateway	2	m1.medium		$0.066	0.132	$600	Heavy	$3,100

This sample design makes use of a combination of two Citrix Secure Gateway and Citrix Web Interface instances to provide access, one per AWS availability zone. It uses one dedicated zone data collector and XML broker, one per Amazon Availability Zone, as well as two core XenApp Server instances available 24 hours in session-only mode.

The above shown spreadsheet has been made available as a tool to assist you in calculating your own estimated costs for running you XenApp farm on Amazon Web Services.

It can be downloaded from the Citrix Developer Network community page for Amazon Web Services. (http://community.citrix.com/display/cdn/Amazon)
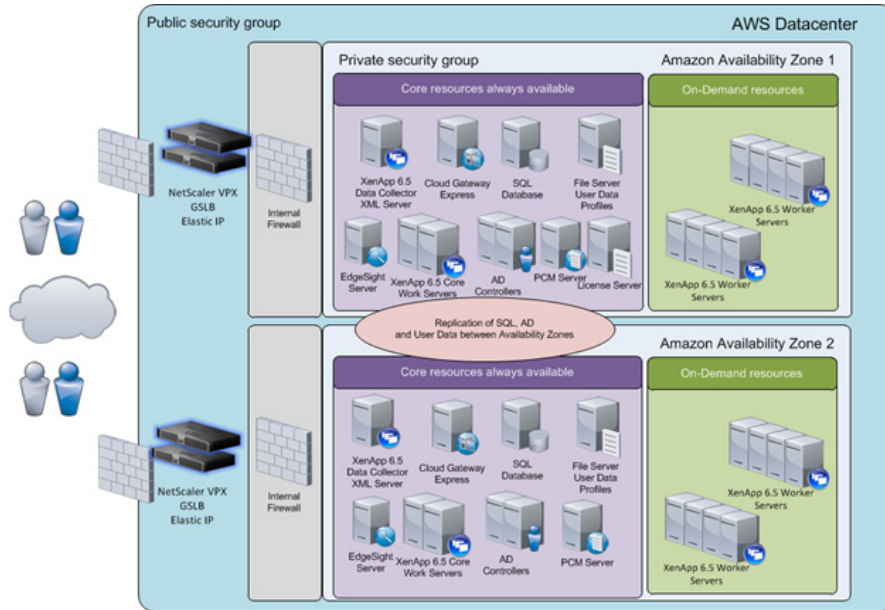
Figure 3 XenApp Farm on AWS for 10,000 hosted shared desktops

Table 2 XenApp on AWS 10,000 desktops costing information
Architecture up to 10,000 users

		Costing Totals	Daily	Monthly	
Cost per user per year	$69.74	XenApp Servers Elastic Capacity	$982.55	$29,885.84	Annual
Cost per user per month	$5.82	Infrastructure Services @24 hrs	$285.96	$8,697.95	$358,630.02
Cost per user per day	$0.20	EBS Storage cost	$350.72	$10,667.80	$104,375.40
Cost per user per hour	$0.01	Network Data transfer cost	$42.00	$840.00	$128,013.60
Supported Logons per minute	720	Reserve Charges	$257.73	$7,839.17	$10,080.00
Minutes to logon all users	14	Total Recurring Cost	$1,925.04	$58,115.75	$94,070.00
					$697,389.02

XenApp Servers		Instance Type	Cost per hr	Extension			
Number of users	10000					Reserve Charge	
Users per server	150						
Uptime per day (hours)	14						
Number of XA Servers require	63	cc2.8xlarge	$1.114	70.182			
				3 Year Term - Reserved Instances			$2,710

Infrastructure Services		Instance Type	Cost per hr	Extension			
Number of core XA servers	4	cc2.8xlarge	$0.571	2.284	Reserve Charge	24hrs	Total Reserve
Data Collector	2	c1.xlarge	$0.420	0.84	$7,670	Heavy	$30,680
SQL Server	3	cc1.4xlarge	$0.477	1.431	$3,100	Heavy	$6,200
Cloud Gateway Express	6	m2.xlarge	$0.158	0.948	$6,300	Heavy	$18,900
License Servers	1	c1.xlarge	$0.420	0.42	$1,550	Heavy	$9,300
Active Directory Servers	6	c1.xlarge	$0.420	2.52	$3,100	Heavy	$3,100
EdgeSight Web Server	2	c1.xlarge	$0.420	0.84	$2,000	Heavy	$18,600
Windows File Server	4	c1.xlarge	$0.420	1.68	$2,000	Heavy	$6,200
Power Capacity Manager	2	c1.xlarge	$0.420	0.84	$2,000	Heavy	$12,400
NetScaler VPX	4	m2.xlarge	$0.158	0.632	$1,030	Heavy	$6,200
						Heavy	$6,200

This sample design makes use of a combination of Citrix NetScaler® and Citrix Access Gateway™ Enterprise VPX and Cloud Gateway Express instances to provide access, two per Amazon Availability Zone. It uses two dedicated zone data collectors and XML brokers, one per Amazon Availability Zone, as well as four core XenApp server instances available 24 hours in session-only mode. An SQL DB Server is set up in redundant mode using SQL Mirroring.

Database mirroring can be used to maintain a single standby database, or mirror database, for a corresponding production database that is referred to as the principal database. Database mirroring runs with either synchronous operation in high-safety mode, or asynchronous operation in high-performance mode. In high-safety mode with automatic failover, which is recommended for XenApp, a third server instance, known as a witness, is required. This enables the mirror server to act as a hot standby server. Failover from the principal database to the mirror database happens automatically and typically takes several seconds.

Scalability testing in a multi-user environment

Several important factors need to be taken into consideration when determining the number of users an instance can support. Sizing a XenApp server Instance depends on the following criteria:

• Specifications of the instance type, such as CPU, memory, disk and network capacity

• Applications, their requirements and how they interact with the system

• Degree of user activity

Before any tests are executed, instance resource limits and response time thresholds need to be set in order to determine the number of users a particular instance can support, for example:

• Eighty percent CPU usage

• Seventy percent memory usage, up to the point where paging begins to saturate the disk

• Acceptable user response times (≤ 4000ms)

These limits should be set to accommodate utilization peaks and future growth; optimal user density of an instance does not equate to its maximum capacity. If you size a server based solely on its maximum capacity, you can expect problems, such as instance instability and poor user response times. Thus, in each performance test, consider the relationship between instance performance and the user experience.

For example, high CPU utilization can be the result of a low memory condition where the system is swapping applications from memory to disk, causing a large number of disk interrupts. Without collecting data for all of the machine's major subsystems, we would easily conclude that the bottleneck was the CPU, when it was a memory bottleneck all along.

Please reference the chart in Appendix A for performance counters for all of the major subsystems (CPU, memory, disk and network) that should be monitored in these types of scalability tests.

Test results

The following section gives a detailed performance analysis for the tests used to determine the scalability impact of virtual processors (vCPU) and the impact of multiple virtual machines executing concurrently on the same physical server. The Login VSI performance tool used to conduct the scalability tests provides a mechanism to automatically collect user experience information (VSIIndex_avg). In order to determine the number of users a configuration can support, this user experience data is overlaid with the performance data of the subsystem that experienced a performance bottleneck. The point where user experience and the subsystem bottleneck intersect determines the number of users that the configuration can support. For convenience, a chart with all of the subsystem data is displayed for each test.

The guidelines presented in this section should be used as a reference, and are only valid if your workloads and applications are similar to those executed in our tests. As always, we recommend executing your own scalability tests, utilizing your server and applications when making your final sizing decisions.

Test Setup and configurations

In this project, we tested the different available Amazon EC2 instances as single instances and we also tested them as part of a XenApp 6.5 Farm hosting 1,000 users to illustrate the linear scalability.

AWS instance test configuration for single instance scalability
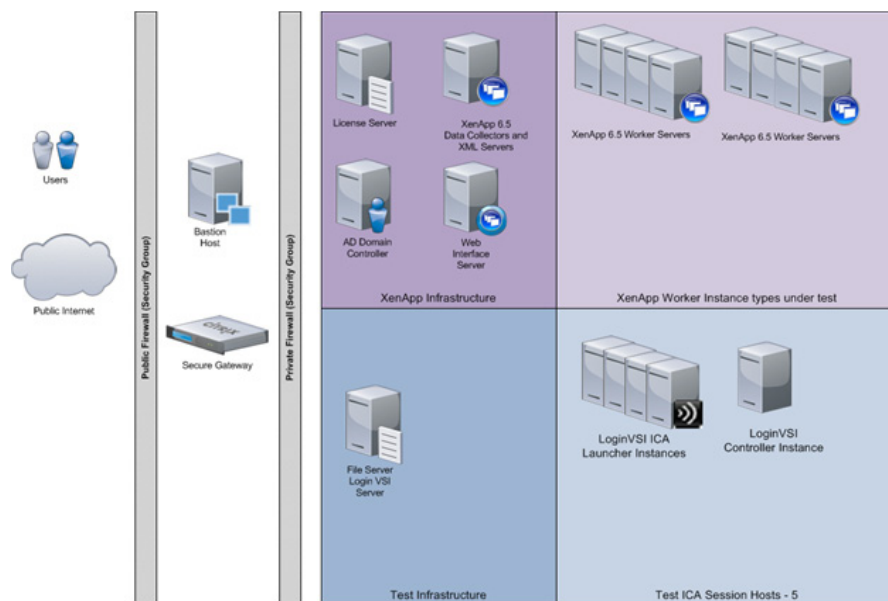
AWS Instances

- 1 x domain controller also functioning as DNS server

- 1 x dedicated XenApp DC/XML Broker

- 1 x Web Interface

- 1 x VSI controller

- 5 x VSI launchers

- AWS Security Groups

  - Private security group

  - Public security group

- XenApp Instances are configured with a single EBS boot volume and 1-2 local-instance storage volumes, depending on the instance type. The local instance volumes are configured to hold volatile data such as the pagefile and user profiles

Software components

• Standard Windows 2008 R2 AMIs 64-bit

• XenApp 6.5 with latest public Hotfixes

• Web Interface 5.4

• Citrix Receiver 3.1

• Office 2010

• LoginVSI 3.5

XenApp 6.5 configuration for 1,000 user scalability
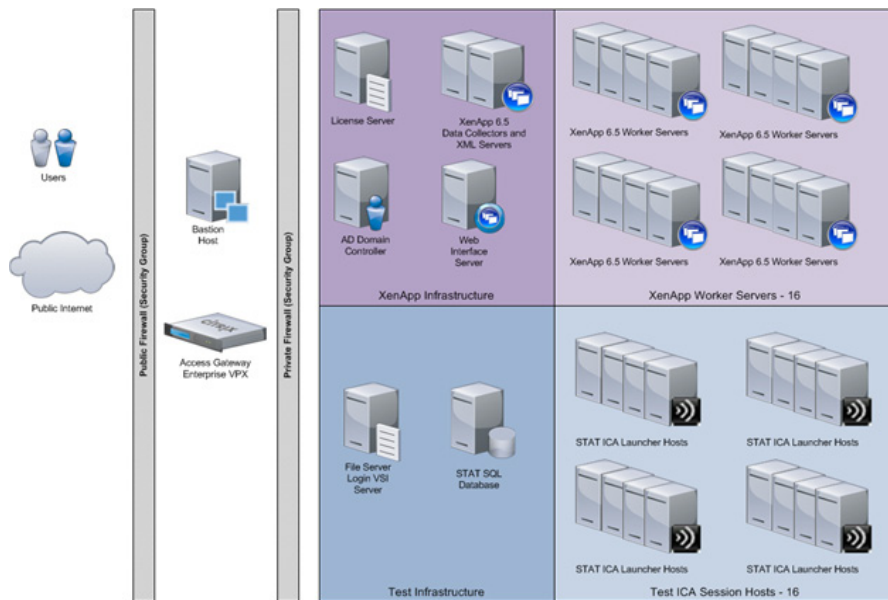
AWS Components

- 1 x domain controller, also functioning as DNS server

- 1 x dedicated XenApp DC/XML Broker

- 1 x Web Interface

- 16 x VSI launchers

- AWS Security Groups

  - Private security group

  - Public security group

- EBS Volumes for XenApp server instances

  - 1 x 30GB EBS boot volume

  - 5 x 10GB EBS volumes configured as a single striping set
    XenApp servers can generate very high amounts of IOPS. It is therefore advised
    to spread the disk I/O over multiple EBS volumes. This is to ensure that the
    configured storage for the XenApp server instances can provide the required
    amount of IOPS. Amazon EBS is namely a shared infrastructure. Using multiple
    EBS volumes will reduce the impact of other customer workloads.

Software components

- Standard Windows 2008R2 AMIs 64-bit

- XenApp 6.5 with latest public Hotfixes

- Web Interface 5.4

- Citrix Receiver 3.1

- Office 2010

- LoginVSI 3.5

Testing methodology

Overview

The objective of the test was to demonstrate XenApp 6.5 scaling capabilities and to understand the Amazon EC2 and EBS utilization.

Through conducting the same series of single-server scalability tests against the different instance types, we were able to run consistent and comparable tests.

Load generation

Within each test environment, load generators were utilized to put demand on the system to simulate multiple users accessing the XenApp 6.5 environment and executing a typical end-user workflow. To generate load within the environment, an auxiliary software application was required to generate the end-user connection to the XenApp environment; to provide unique user credentials, to initiate the workload; and to evaluate the end user experience.

In the hosted shared desktop test environment, sessions launchers were used to simulate multiple users making a direct connection to XenApp 6.5 via a Citrix HDX protocol connection.

User workload simulation – Login VSI from Login Consultants

One of the most critical factors of validating a XenApp deployment is identifying a real-world user workload that is easy for customers to replicate and standardize across platforms in order to allow customers to realistically test the impact of a variety of worker tasks. To accurately represent a real-world user workload, a third-party tool from Login Consultants was used throughout the Hosted Shared Desktop testing.

The tool has the benefit of taking measurements of the in-session response time, which provides an objective way to measure the expected user experience for an individual desktop throughout large-scale testing, including login storms.

The Virtual Session Indexer, Login Consultants' Login VSI 3.5 methodology, designed for benchmarking Server Based Computing (SBC) and Virtual Desktop Infrastructure (VDI) environments is completely platform and protocol independent and hence allows customers to easily replicate the testing results in their environment.

Login VSI calculates an index based on the amount of simultaneous sessions that can be run on a single machine.

Login VSI simulates a medium workload user, also known as a knowledge worker, running generic applications such as: Microsoft Office 2007 or 2010, Internet Explorer 8 including a Flash video applet and Adobe Acrobat Reader. Note: For the purposes of this test, applications were installed locally, not streamed or hosted on other XenApp servers.

Like real users, the scripted Login VSI session will leave multiple applications open at the same time. The medium workload is the default workload in Login VSI and was used for this testing. This emulated a knowledge worker using Office, IE, printing and PDF.

- Once a session has been started, the medium workload will repeat every 12 minutes

- During each loop, the response time is measured every two minutes

- The medium workload opens up to five apps simultaneously

- The type rate is 160ms for each character

- Approximately two minutes of idle time is included to simulate real-world users

Each loop will open and use:

- Outlook 2007/2010: browse 10 messages

- Internet Explorer: one instance is left open (BBC.co.uk); one instance is browsed to Wired.com, Lonelyplanet.com and heavy 480p Flash application gettheglass.com

- Word 2007/2010: one instance to measure response time; one instance to review and edit document

- Bullzip PDF Printer and Acrobat Reader: the word document is printed and reviewed to PDF

- Excel 2007/2010: a very large randomized sheet is opened

- PowerPoint 2007/2010: a presentation is reviewed and edited

- 7-zip: using the command-line version, the output of the session is zipped

A graphical representation of the medium workload is shown in Figure 55.

Figure 4 Graphical representation of medium workload
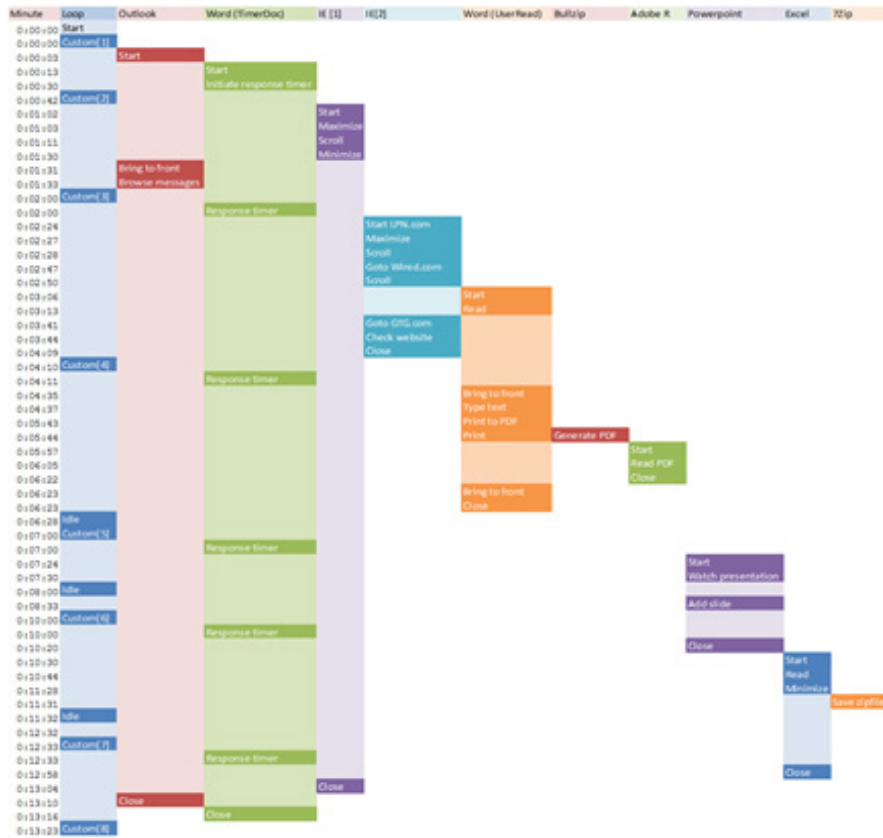
You can obtain additional information on Login VSI from http://www.loginvsi.com.

Testing procedure

The following protocol was used for each test cycle in this study to ensure consistent results.

Pre-test setup for single-instance and complete farm testing

All instances under this test were shut down utilizing the AWS Console. All launchers for the test were shut down. They were then restarted until the required number of launchers were running with the Login VSI Agent at a "waiting for test to start" state.

Success criteria

There were multiple metrics that were captured during each test run, but the success criteria for considering a single test run as pass or fail was based on the key metric, VSI Max. The Login VSI Max evaluates the user response time during an increased user load and assesses the successful start-to-finish execution of all the initiated virtual desktop sessions.

Login VSI Max

VSI Max represents the maximum number of users the environment can handle before serious performance degradation occurs. VSI Max is calculated based on the response times of individual users as indicated during the workload execution. The user response time has a threshold of 4000ms and all users' response times are expected to be less than 4000ms in order to assume that the user interaction with the virtual desktop is at a functional level. VSI Max is reached when the response times reach or exceed 4000ms for six consecutive occurrences.

If VSI Max is reached, that indicates the point at which the user experience has significantly degraded. The response time is generally an indicator of the host CPU resources, but this specific method of analyzing the user experience provides an objective method of comparison that can be aligned to host CPU performance.

Note: In the prior version of Login VSI, the threshold for response time was 2000ms. The workloads and the analysis have been upgraded in Login VSI 3 to make the testing more aligned to real-world use. In the medium workload in Login VSI 3.0, a CPU-intensive 480p flash movie is incorporated in each test loop. In general, the redesigned workload would result in an approximate 20 percent decrease in the number of users passing the test, versus Login VSI 2.0 on the same server and storage hardware.

Calculating VSI Max

Typically, when a simulated Login VSI user is logged on, the desktop workload is scripted in a 12 -14 minute loop. After the loop is finished, it will restart automatically. Within each loop, the response time of seven specific operations is measured in a regular interval: six times within each loop.

The seven operations from which the response times are measured are:

- Copy new document from the document pool in the home drive

    - This operation will refresh a new document to be used for measuring the response time. This activity is mostly a file-system operation

- Starting Microsoft Word with a document

    - This operation will measure the responsiveness of the operating system (OS) and the file system. Microsoft Word is started and loaded into memory; also the new document is automatically loaded into Microsoft Word. When the disk I/O is extensive or even saturated, this will impact the File Open dialogue considerably

- Starting the File Open dialogue

    - A small part of this operation is handled by Word and a large part by the operating system. The File-Open dialogue uses generic subsystems and interface components of the OS. The OS provides the contents of this dialogue

- Starting Notepad

    - This operation is handled by the OS (loading and initiating Notepad.exe) and by the Notepad.exe itself through execution. This operation seems instant from an end-user's point of view

- Starting the Print dialogue

  - A large part of this operation is handled by the OS subsystems, as the Print dialogue is provided by the OS. This dialogue loads the print subsystem and the drivers of the selected printer. As a result, this dialogue is also dependent on disk performance

- Starting the Search and Replace dialogue

  - This operation is handled within the application completely; the presentation of the dialogue is almost instant. Serious bottlenecks on the application level will impact the speed of this dialogue

- Compress the document into a zip file with 7-zip command line

  - This operation is handled by the command-line version of 7-zip. The compression will very briefly spike CPU and disk I/O

These measured operations with Login VSI do hit considerably different subsystems, such as CPU (user and kernel), memory, disk, the OS in general, the application itself, print, GDI, etc. These operations are specifically short by nature. When such operations are consistently long, the system is saturated because of excessive queuing on any kind of resource. As a result, the average response times will then escalate. This effect is clearly visible to end-users. When such operations consistently consume multiple seconds, the user will regard the system as slow and unresponsive.

VSI Max Classic

VSI Max Classic is based on the previous versions of Login VSI, and is achieved when the average Login VSI response time is higher than a fixed threshold of 4000ms. This method proves to be reliable when no anti-virus or application virtualization is used.

To calculate the response times, the seven activities listed in the previous section are totaled. To balance these measurements, they are weighted before they are summed. Without weighting individual response times before they are totaled, one specific measurement (out of seven) could dominate the results.

Within VSI Max Classic, two measurements are weighted before they are added to the total VSI Max response time:

1. 'Starting Microsoft Word with a document is divided by two (50%)

2. 'Starting the Search-and-Replace dialogue is multiplied by five (500%)

A sample of the VSI Max Classic response time calculation is displayed in Table 16.

Table 3 VSI Max Classic response time calculation

Activity (RowName)	Result (ms)	Weight (%)	Weighted Result (ms)
Refresh document (RFS)	160	100%	160
Start Word with new doc (LOAD)	1400	50%	700
File Open Dialogue (OPEN)	350	100%	350
Start Notepad (NOTEPAD)	50	100%	50
Print Dialogue (PRINT)	220	100%	220
Replace Dialogue (FIND)	10	500%	50
Zip documents (ZIP)	130	100%	130
VSImax Classic Response Time			**1660**

Then the average VSI Max response time is calculated, based on the amount of active Login VSI users logged on to the system. When the average VSI Max response times are consistently higher than the default threshold of 4000ms, VSI Max is achieved.

In practice, however; tests have shown a substantial increase of application response time when antivirus and/or application virtualization is used. The baseline response time is typically around 1400–1800ms without application virtualization or antivirus. However, when anti-virus or application virtualization is used, the baseline response time varies between 2500–3500ms.

When the baseline response time is already so high, the VSI Max threshold of 4000ms is too easily reached. VSI Max Classic will report having reached a maximum, long before system resources like CPU, memory or disk are actually saturated.

It was therefore decided to further optimize VSI Max calculation. As of Login VSI 3.0 and later, VSI Max Dynamic is introduced, to be able to support wildly varying baseline response times when anti-virus and/or application virtualization is used.

VSI Max Classic was not used in this study.

VSI Max Dynamic

Similar to VSI Max Classic, VSI Max Dynamic is calculated when the response times are consistently above a certain threshold. However, this threshold is now dynamically calculated on the baseline response time of the test.

Five individual measurements are weighted to better support this approach:

• Copy new doc from the document pool in the home drive: 100%

• Microsoft Word with a document: 33.3%

• Starting the File Open dialogue: 100%

• Starting Notepad: 300%

• Starting the Print dialogue: 200%

• Starting the Search and Replace dialogue: 400%

• Compress the document into a zip file with 7-zip command line: 200%

A sample of the VSI Max Dynamic response time calculation is displayed in Table 17.

Table 4 VSI Max Dynamic Response Time calculation

Activity (RowName)	Result (ms)	Weight (%)	Weighted Result (ms)
Refresh document (RFS)	160	100%	160
Start Word with new doc (LOAD)	1400	33.3%	467
File Open Dialogue (OPEN)	350	100%	350
Start Notepad (NOTEPAD)	50	300%	150
Print Dialogue (PRINT)	220	200%	440
Replace Dialogue (FIND)	10	400%	40
Zip documents (ZIP)	130	200%	230
VSImax Dynamic Response Time	**1837**		

Then the average VSI Max response time is calculated, based on the amount of active Login VSI users logged on to the system. For this, the average VSI Max response times need to be consistently higher than a dynamically calculated threshold.

To determine this dynamic threshold, first the average baseline response time is calculated. This is done by averaging the baseline response time of the first 15 Login VSI users on the system.

The formula for the dynamic threshold is: Average Baseline Response Time x 125% + 3000. As a result, when the baseline response time is 1800, the VSI Max threshold will now be 1800 x 125% + 3000 = 5250ms.

The baseline response time can wildly vary per vendor and streaming strategy, especially when application virtualization is used. Therefore, it is recommended that you use VSI Max Dynamic when comparisons are made with application virtualization or anti-virus agents. The resulting VSI Max Dynamic scores are aligned again with saturation on a CPU, memory or disk level and also when the baseline response times are relatively high.

Determining VSI Max

The Login VSI analyzer will automatically identify the VSI Max. In the example below, the VSI Max is 98. The analyzer will automatically determine "stuck sessions" and correct the final VSI Max score.

• Vertical axis: response time in milliseconds

• Horizontal axis: total active sessions

Figure 5 Sample Login VSI Analyzer Graphic Output
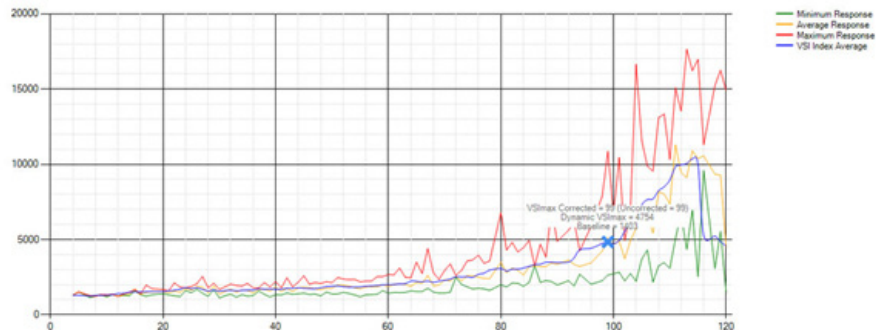
- Red line: maximum response (worst response time of an individual measurement within a single session)

- Orange line: average response time within for each level of active sessions

- Blue line: the VSI Max average

- Green line: minimum response (best response time of an individual measurement within a single session)

In our tests, the total number of users in the test run had to login, become active and run at least one test loop and log out automatically without reaching the VSI Max to be considered a success.

The Citrix AppCenter will be monitored throughout the steady state to ensure that:

- All running sessions report In Use throughout the steady state

- No sessions move to Unregistered or Available state at any time during Steady state

Within 20 minutes of the end of the test, all sessions on all launchers must have logged out automatically and the Login VSI Agent must have shut down.

Test results

The purpose of this testing is to provide the data needed to validate Citrix XenApp 6.5 Hosted Shared Desktop FlexCast™ model on Amazon EC2 to virtualize Microsoft Windows 2008R2 desktops.

The information contained in this section provides data points that a customer may reference while designing their own implementations. These validation results are an example of what is possible under the specific environment conditions outlined in this paper, and do not represent the full characterization of XenApp on Amazon EC2.

The primary success criteria metrics are provided to validate the overall success of the test cycle as an output chart from Login Consultants' VSI Analyzer Professional Edition, VSI Max Dynamic for the medium workload (with Flash).

Additional graphs detailing the CPU and memory utilization during peak session load are also presented. Given adequate storage capability, the limiting factor in the testing was CPU utilization.

Single-instance validation
Cluster Compute Instances

Instances of this family provide proportionally high CPU and memory resources with increased network performance and are well suited for high-performance computing (HPC) applications but also for Hosted Shared Desktop applications and other demanding network-bound applications. These instance types are the most suited instance types for the XenApp worker, due to their ability to host large amounts of XenApp Hosted Shared Desktop sessions, using instances configured with respectively 16 or 32 vCPUs, combined with two high-performance local-instance disks.

Cluster Compute Eight Extra Large Instance – cc2.8xlarge

This section details the results from the XenApp Hosted Shared Desktop single instance validation testing on Cluster Compute Eight Extra Large Instance with the API.name cc2.8xlarge.

A Cluster Compute, Eight Extra Large instance comes with 60.5 GB of memory, 88 EC2 Compute Units (2 x Intel Xeon E5-2670, eight-core Sandy Bridge architecture), 1690 GB of local-instance storage, and very high I/O performance using a 10GbE interconnect.

The Cluster Compute Eight Extra Large Instance has been configured with the OS and applications on the EBS boot volume, and the pagefile and EdgeSight DB on one of the local-instance storage volumes, and all user profiles on the other local instance storage volume.

Configured in this way, it enabled for 150 Hosted Shared Desktops with the Login VSI medium workload. The limiting factor was the EBS boot volume, hosting the OS and applications.

Figure 6 150 desktop sessions on cc2.8xlarge Windows Server 2008 R2 SP1 below 4000 ms
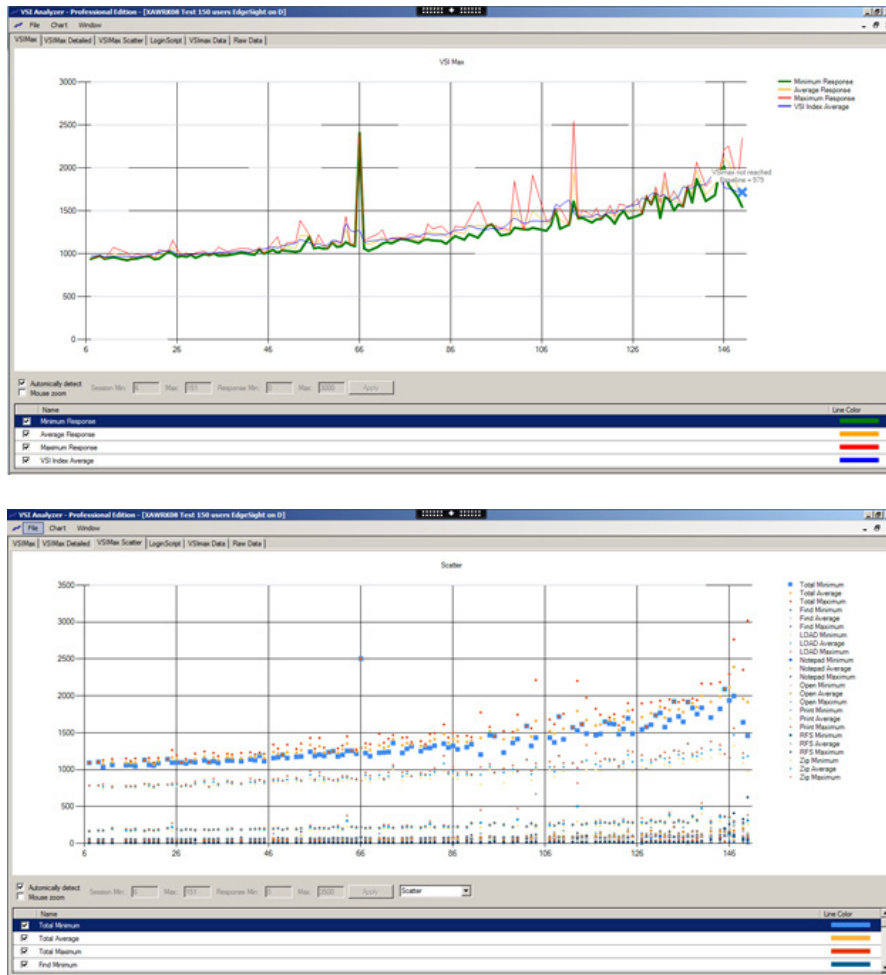
The following graphs detail CPU, memory, disk and network performance on the Amazon EC2 cc2.8xlarge instance. Disk IOPS for the EBS boot volume was the limiting factor, as you can see from the Disk Transfer/Sec graph for the C: volume.
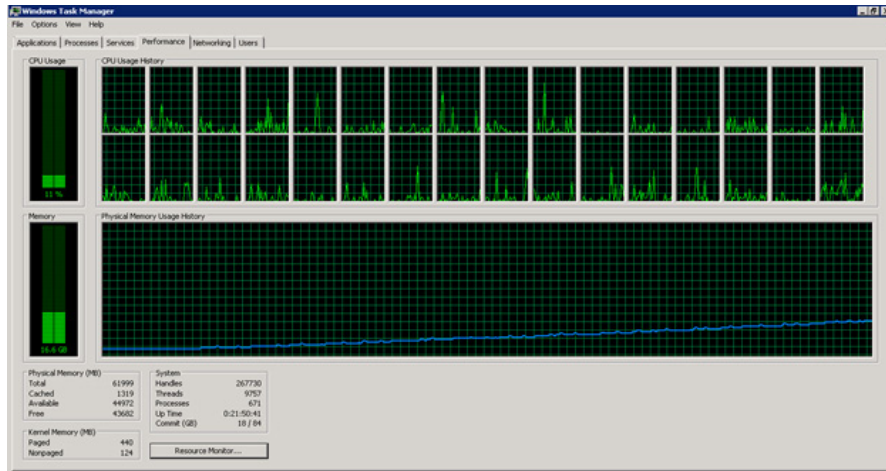
Figure 7 150 User Single cc2.8xlarge instance CPU utilization all phases
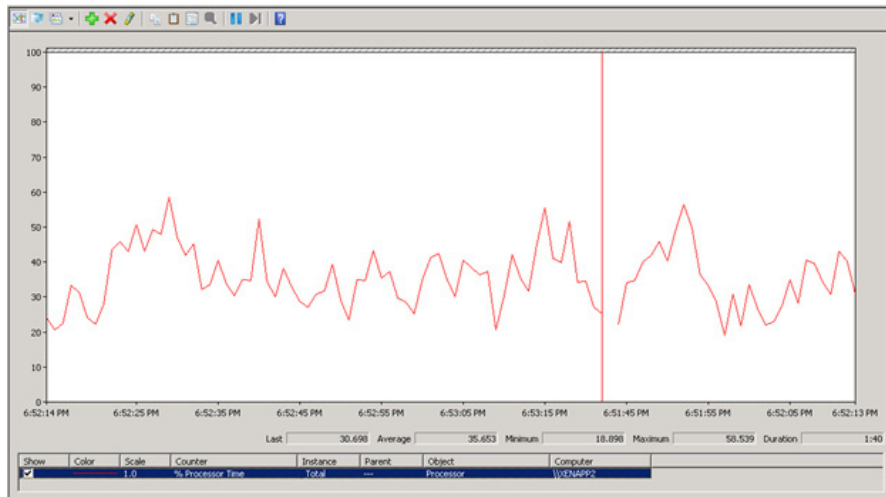
Figure 8 150 users CloudWatch CPU utilization of cc2.8xlarge instance
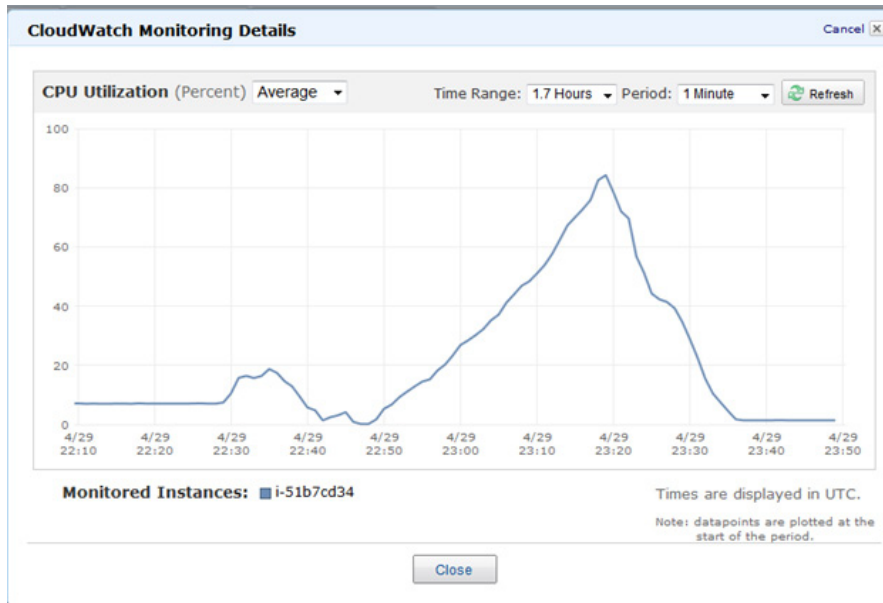
Figure 9 150 User Single cc2.8xlarge instance available memory
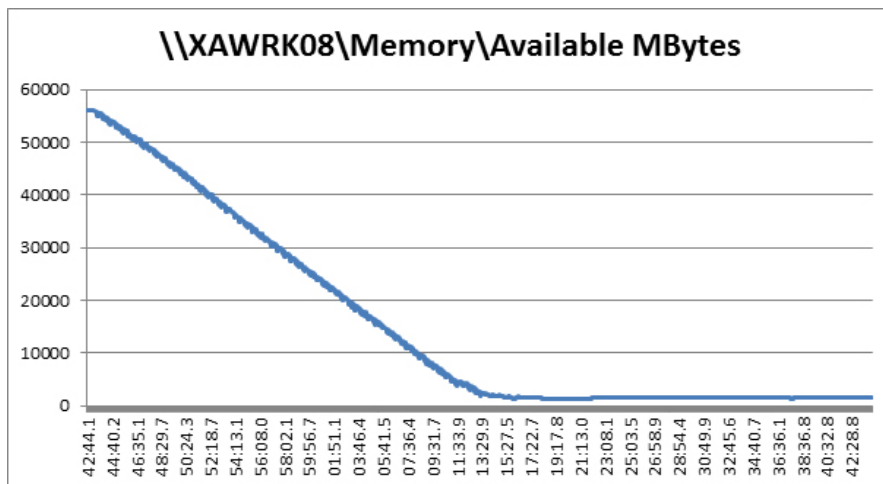
Figure 10 Single cc2.8xlarge instance Mbits received per second

Figure 11 Single cc2.8xlarge instance average disk queue length for EBS boot volume
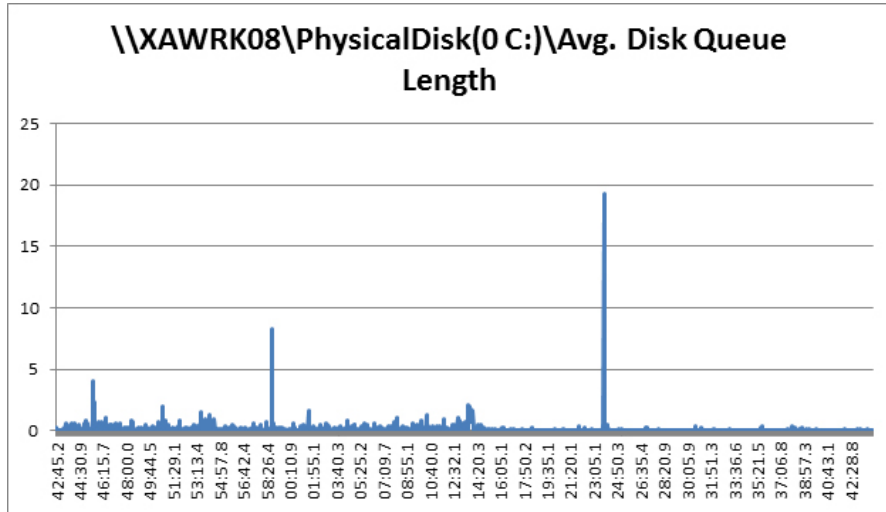
Figure 12 Single cc2.8xlarge instance disk IOPS for EBS boot volume

—

—
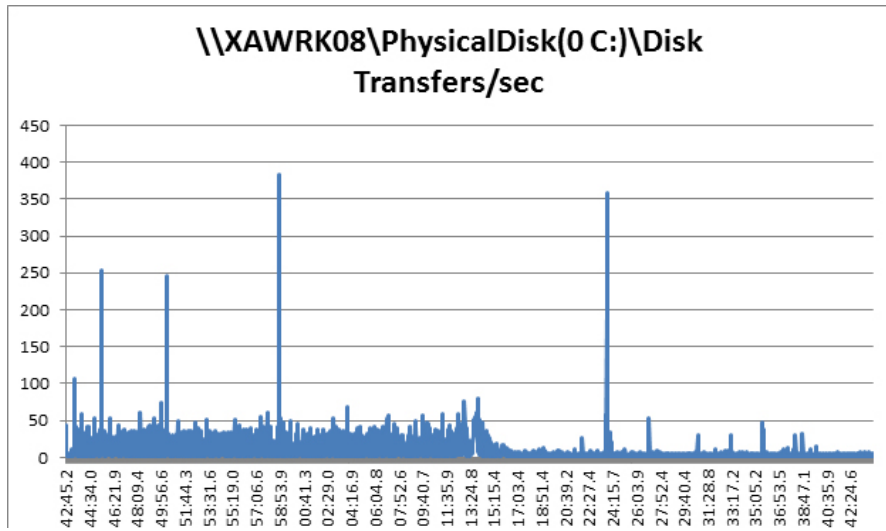
Figure 13 Single cc2.8xlarge instance average disk queue length for local instance Pagefile disk
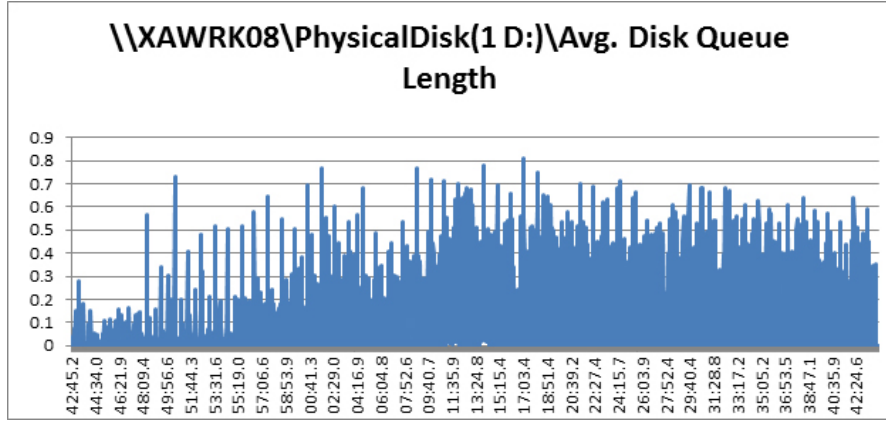
Figure 14 Single cc2.8xlarge instance disk IOPS for local instance pagefile disk
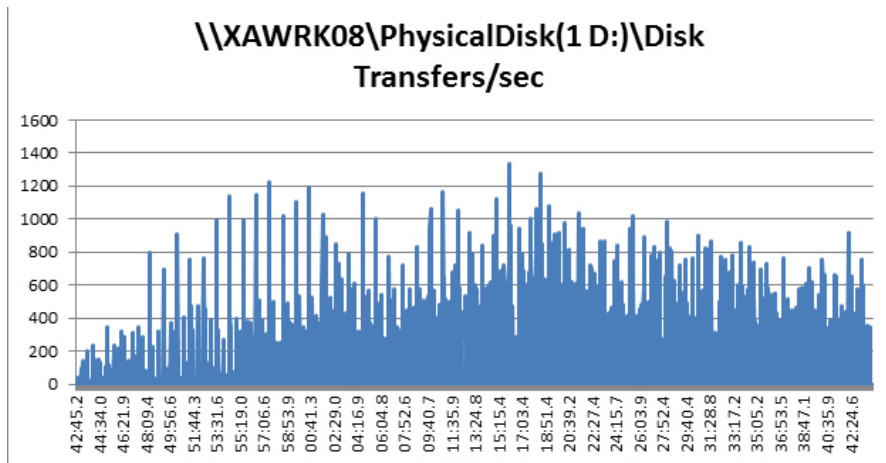
Figure 15 Single cc2.8xlarge instance average disk queue length for local instance user profile disk
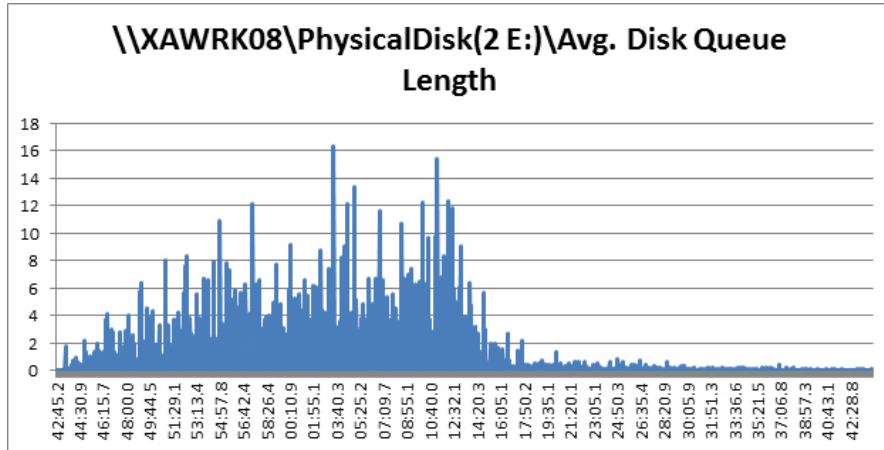
Figure 16 Single cc2.8xlarge instance Disk IOPS for local instance user profile disk
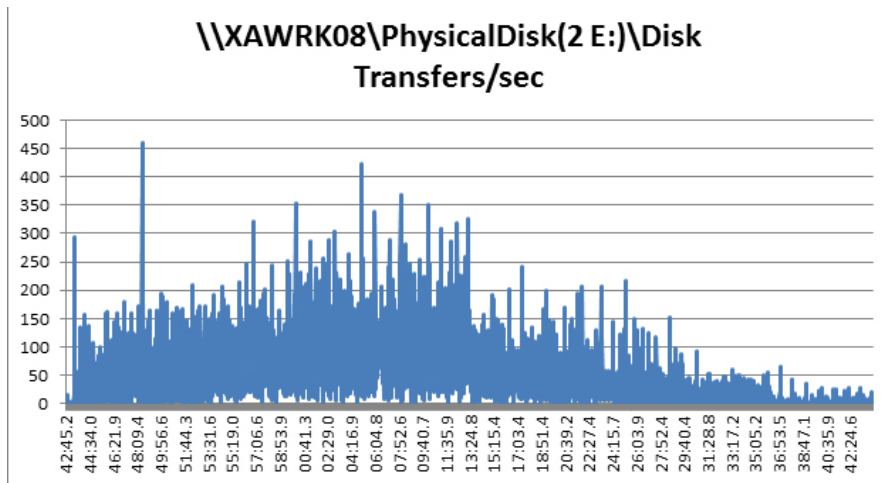
Cluster Compute Quadruple Extra Large Instance – cc1.4xlarge

This section details the results from the XenApp Hosted Shared Desktop single instance validation testing on Cluster Compute Quadruple Extra Large Instance with the API.name cc1.4xlarge.

A Cluster Compute Quadruple Extra Large instance comes with 23 GB of memory, 33.5 EC2 Compute (2 x Intel Xeon X5570, quad-core "Nehalem" architecture), 1690 GB of local instance storage, and very high I/O performance using a 10GbE interconnect.

The limiting factor for the Cluster Compute Quadruple Extra Large instance is the 23GB of memory, still however hosting 80 Hosted Shared Desktop sessions.

Figure 17 80 Desktop Sessions on cc1.4xlarge Windows Server 2008 R2 SP1 below 4000 ms

High-CPU Instances

Instances of this family have proportionally more CPU resources than memory (RAM) and are well suited for compute-intensive applications. Not all of these instances are suited for a XenApp worker. The High-CPU Medium Instance, for example, only comes with 1.7GB, making it therefore unusable for a XenApp worker.

High-CPU Extra Large Instance – c1.xlarge

This section details the results from the XenApp Hosted Shared Desktop single-instance validation testing on High-CPU Extra Large Instance with the API.name c1.xlarge.

A High-CPU Extra Large instance comes with 7 GB of memory, 20 EC2 Compute Units (8 virtual cores with 2.5 EC2 Compute Units each), 1690 GB of local instance storage, and high I/O performance.

Figure 18 22 Desktop Sessions on c1.xlarge Windows Server 2008 R2 SP1 below 4000 ms

High-Memory Instances

Instances of this family offer large memory sizes for high-throughput applications, including database and memory-caching applications. These instances are reasonably suited for a XenApp worker. They offer a reasonable good balance between CPU and memory for a XenApp worker.

High-Memory Quadruple Extra Large Instance – m2.4xlarge

This section details the results from the XenApp Hosted Shared Desktop single instance validation testing on High-Memory Quadruple Extra Large instance with the API.name m2.4xlarge.

A High-Memory Quadruple Extra Large instance comes with 68.4 GB of memory, 26 EC2 Compute Units (8 virtual cores with 3.25 EC2 Compute Units each), 1690 GB of local instance storage, and high I/O performance.

Figure 19 65 Desktop Sessions on c1.xlarge Windows Server 2008 R2 SP1 below 4000 ms
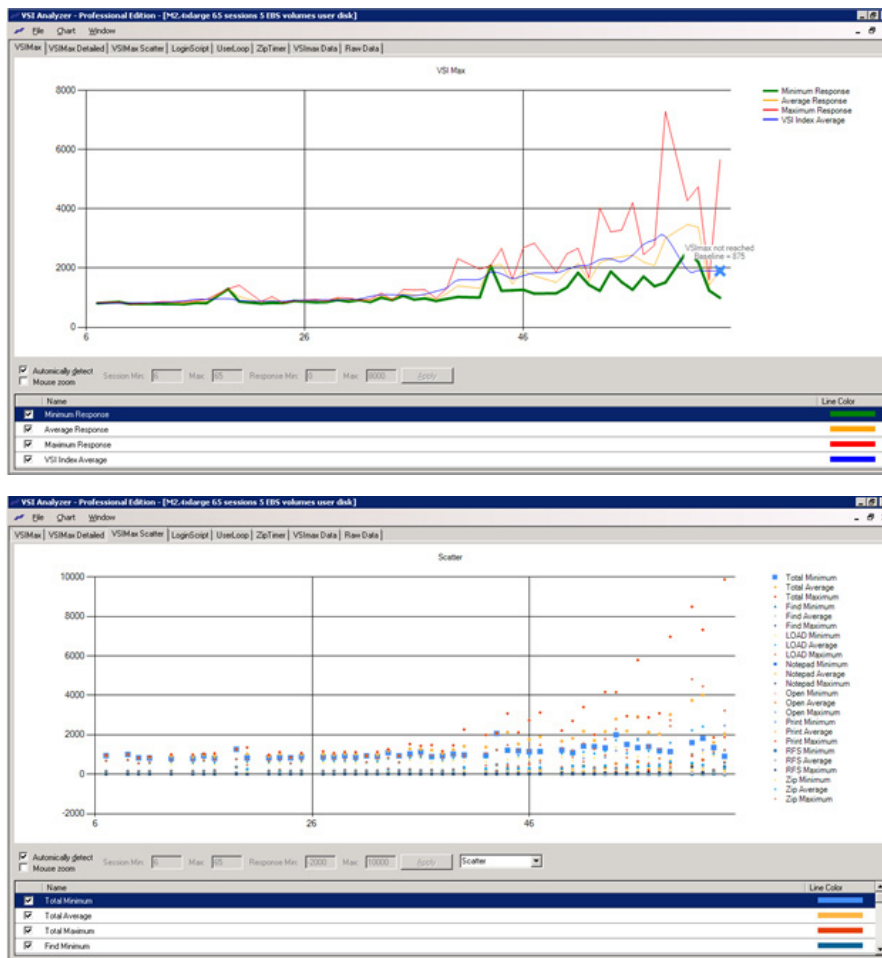
High-Memory Double Extra Large Instance – m2.2xlarge

This section details the results from the XenApp Hosted Shared Desktop single instance validation testing on High-Memory Double Extra Large instance with the API.name m2.2xlarge.

A High-Memory Double Extra Large instance comes with 34.2 GB of memory, 13 EC2 Compute Units (4 virtual cores with 3.25 EC2 Compute Units each), 850 GB of local instance storage, and high I/O performance.

Figure 20 33 desktop sessions on m2.2xlarge Windows Server 2008 R2 SP1 below 4000 ms
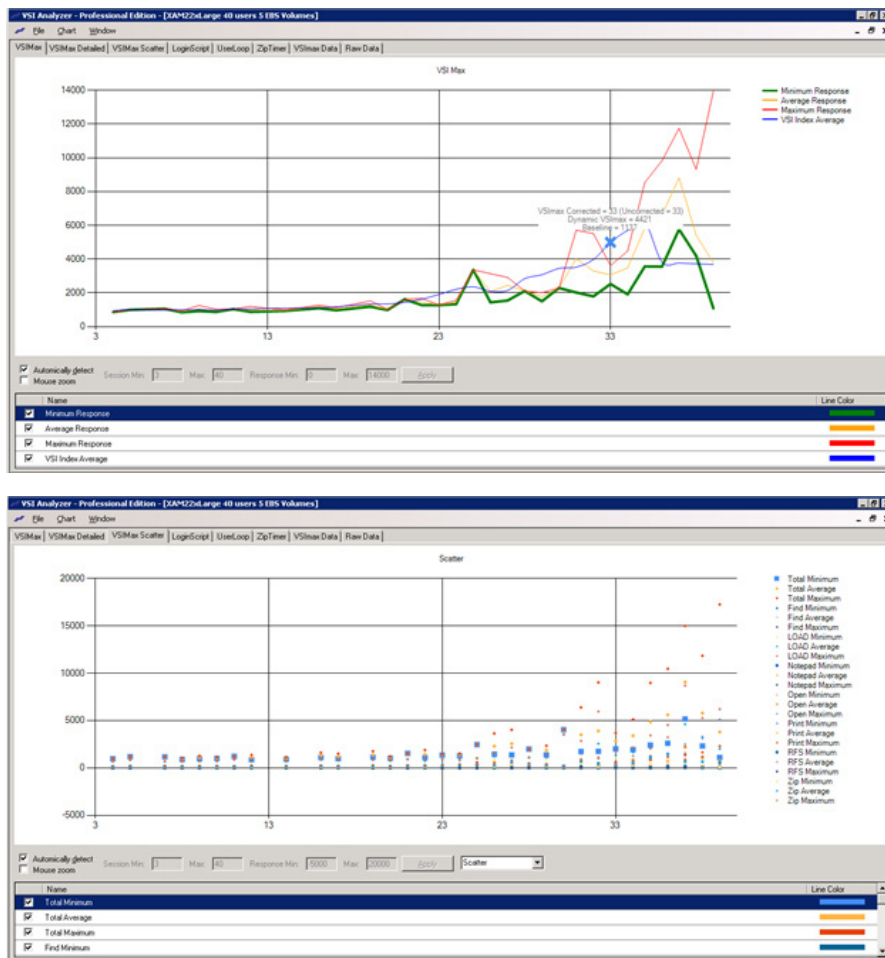
Multi-instance validation – 1,000 users

This section details the results from the XenApp Hosted Shared Desktop sixteen instance validation testing. It illustrates linear scalability from one instance with 67 Users to 15 instances with 1,000 users.

Each instance was configured as a m2.4xlarge instance with a single EBS boot volume using the Login VSI light user workload, since the purpose of this specific test was to demonstrate the feasibility of a large amount of XenApp server Worker Instances as part of a XenApp Workgroup on AWS.

For this test, we used an additional Citrix internal instrument named STAT that is able to control the many Login VSI Launchers from a single central console and also collects performance data from all involved components. This allows us to graph the progress of HDX sessions, logging in successfully to the XenApp Farm.

Additional graphs detailing the CPU and utilization during peak-session load are also presented. Given adequate storage capability, the limiting factor in the testing was CPU utilization. Each of the 16, m2.4xlarge instances performance charts are essentially identical for the multi-instance runs. We are including AWS CloudWatch data on all of the 16 instances, for CPU, disk and network utilization.

Figure 21 1000 Hosted Shared Desktop Session on 16 AWS EC2 m2.4xlarge instances

Figure 22 67 Hosted Shared Desktop Session per AWS EC2 m2.4xlarge instances

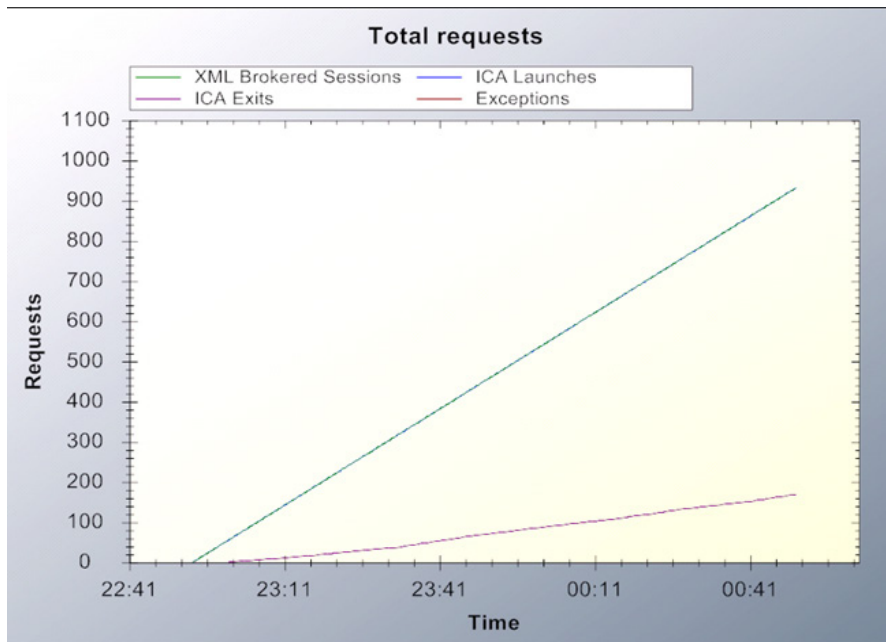Figure 23 1000 Successful Hosted Shared Desktops on AWS

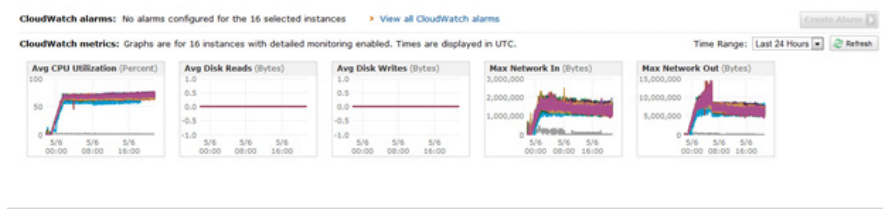Figure 24 CloudWatch Metrics for 16 m2.4xlarge instances

Figure 25 CloudWatch metrics for all 16 EBS volumes

These CloudWatch graphs show a total of nearly 200 IOPS for each server. Since each server is hosting 67 sessions, which leaves us with an average of three IOPS per user for the Login VSI light user workload used in this specific test.

Scalability considerations and guidelines

There are many factors to consider when you begin to deploy instances for a XenApp 6.5 environment on AWS.

In this section, we give some guidance on configuring the AWS instances.

AWS Instance configuration

As our results indicate, we have proven linear scalability in the Amazon EC2 Reference Architecture as tested.

AWS Storage best practices

Storage optimization for deploying XenApp Hosted Shared Desktop solutions on Amazon EBS includes the following:

• Use EBS root volumes to boot hold OS and XenApp installation

• A fully loaded XenApp server instance can easily reach 200+ IOPS, which for a larger XenApp farm would become a very costly exercise for I/Os related to volatile data–especially given that Amazon charges $0.10 per million IOPS per month

• Use, therefore, instance local storage equivalent to a write-cache in a PVS solution, which would contain:

  • Page file

  • EdgeSight Database

  • User profiles

  • Application streaming cache

Here is an example of how to create an AWS instance to be used for XenApp servers with two local instance disks:

Create an Amazon Machine Image (AMI) with local storage enabled, which then would contain the pagefile and user profiles.

The AWS Console does not allow the configuration of the local instance disks, but the EC2 Command-Line tools do and can be found here: http://aws.amazon.com/developertools/739

ec2-run-instances –region us-east- –block-device-mapping "/dev/sdc=ephemeral0" -block-device-mapping "/dev/sdd=ephemeral1" ami-9d10d8f4 -g sg-36e2fb5a -t cc2.8xlarge -k xencloudkey -z us-east-1e –s subnet-940547fc

This command will instantiate a new cc2.8xlarge instance in VPC using a master AMI with a boot disk on EBS and 2 additional drives which are local storage. Note: as of this writing cc2.8xlarge instances are for VPC only available in us-east-e1

• Drive D: for Pagefile and EdgeSight Database, and App-V, XenApp streaming cache

• Drive E: for User Profiles and data

How to change the User Profile directory to point to drive E:

Change the RegKey for HKLM\Software\Microsoft\Windows NT\CurrentVersion\Profilelist\ProfilesDirectory to point to for instance E:\Users instead of %SystemDrive%\Users

Use this in combination with Microsoft GPOs to automatically delete cached roaming profiles after log off. Since the local instance disks are volatile, after a stop and start of the instance, they will no longer contain any data. The EC2Instance service will automatically initialize and format them.

Acknowledgements

Appendix A: performance counters

Counter Name	Recommendation	Description
\Cache\Copy Read Hits %	≥ 98%	Percentage of cache copy read requests that hit the cache; that is, they did not require a disk read to provide access to the page in the cache. A copy read is a file read operation that is satisfied by a memory copy from a page in the cache to the application's buffer.
\LogicalDisk\% Disk Time	≤ 50%	Percentage of elapsed time that the selected disk is busy responding to read or write requests.
\LogicalDisk\% Idle Time	≥ 50%	Percentage of time during the sample interval that the disk was idle.
\LogicalDisk\Current Disk Queue Length	≤ 2	Number of requests outstanding on the disk at the time the performance data is collected.
\Memory\% Committed Bytes In Use	≤ 85%	Ratio of Memory\\Committed Bytes to the Memory\\Commit Limit. Committed memory is the physical memory in use for which space is reserved in the paging file if it needs to be written to disk.
\Memory\Pages Input/sec	≤ 100	Rate at which pages are read from disk to resolve hard page faults.
\Memory\Pages Output/sec	≤ 100	Rate at which pages are written to disk to free space in physical memory.
\Memory\Pool Nonpaged Bytes	≥ 200 MB	Size, in bytes, of the non-paged pool, an area of system memory (physical memory used by the operating system) for objects that cannot be written to disk, but must remain in physical memory as long as they are allocated.
\Memory\Pool Paged Bytes	≥ 200 MB	Size, in bytes, of the paged pool, an area of system memory (physical memory used by the operating system) for objects that can be written to disk when they are not being used
\Network Interface\Bytes Received/sec	≤ Curr. Bandwidth	Rate at which bytes are received over each network adapter, including framing characters
\Network Interface\Bytes Sent/sec	≤ Curr. Bandwidth	Rate at which bytes are sent over each network adapter, including framing characters
\Network Interface\Bytes Total/sec	≤ Curr. Bandwidth	Rate at which bytes are sent and received over each network adapter, including framing characters
\Network Interface\Current Bandwidth	NA	Estimate of the current bandwidth of the network interface in bits per second.
\Paging File\% Usage Peak	≥ 40%	Peak usage of the Page File instance in percent.
\Paging File\% Usage	≥ 40%	Amount of the Page File instance in use in percent.
\Processor\% Processor Time	≤ 65%	Percentage of elapsed time that the processor spends to execute a non-idle thread.
\System\Context Switches/sec	≤ 12000 per Core	The combined rate at which all processors on the computer are switched from one thread to another. Context switches occur when a running thread voluntarily relinquishes the processor, is preempted by a higher priority ready thread, or switches between user-mode and privileged (kernel) mode to use an Executive or subsystem service.
\System\Processor Queue Length	≤ 2 per Core	Number of threads in the processor queue.
\Terminal Services\Active Sessions	NA	Actively connected terminal services sessions.

Appendix B: online resources
Citrix reference documents

Citrix XenDesktop with Login Consultants VSI
http://support.citrix.com/article/CTX125630

Login Consultants VSI References
http://www.loginvsi.com/en/in-practice/references

XenApp Planning Guide - Virtualization Best Practices
http://support.citrix.com/article/CTX129761

The Citrix Service Provider Toolkit
http://community.citrix.com/kits/#/kit/734024

Top 10 Considerations for Delivering Desktops in the Cloud
http://support.citrix.com/article/CTX128899

How to Deliver a Cloud Desktop using XenApp 6
http://support.citrix.com/servlet/KbServlet/download/26821-102-650623/How%20
to%20deliver%20a%20cloud%20desktop%20using%20XenApp%206.pdf

Scaling Big – SaaS and DaaS Deployments for Citrix Service Providers
http://support.citrix.com/article/CTX129106

Microsoft reference documents

Windows Server 2008 R2
http://www.microsoft.com/en-us/server-cloud/windows-server/default.aspx

Active Directory
http://www.microsoft.com/en-us/server-cloud/windows-server/active-directory.aspx

File Services
http://www.microsoft.com/en-us/server-cloud/windows-server/file-and-print-
services.aspx

CITRIX®

Corporate Headquarters	**India Development Center**	**Latin America Headquarters**
Fort Lauderdale, FL, USA	Bangalore, India	Coral Gables, FL, USA
Silicon Valley Headquarters	**Online Division Headquarters**	**UK Development Center**
Santa Clara, CA, USA	Santa Barbara, CA, USA	Chalfont, United Kingdom
EMEA Headquarters	**Pacific Headquarters**	
Schaffhausen, Switzerland	Hong Kong, China	

About Citrix

Citrix Systems, Inc. (NASDAQ:CTXS) is the company transforming how people, businesses and IT work and collaborate in the cloud era. With market-leading cloud, collaboration, networking and virtualization technologies, Citrix powers mobile workstyles and cloud services, making complex enterprise IT simpler and more accessible for 260,000 enterprises. Citrix touches 75 percent of Internet users each day and partners with more than 10,000 companies in 100 countries. Annual revenue in 2011 was $2.21 billion. Learn more at www.citrix.com.