Scalability Considerations for Using the XenApp and XenDesktop Service Local Host Cache Feature with Citrix Cloud Connector

Author: Jahed Iqbal

Overview

The local host cache feature in the XenApp and XenDesktop Service allows connection brokering in a site to continue if there is an outage. An outage happens if the WAN link between the site and the management console fails in a Citrix Cloud environment. In December 2017, we tested the Citrix Cloud Connector machine configuration using the XenApp and XenDesktop Service local host cache feature. The test results provided in this document detail the tested maximums in December 2017. Best practice recommendations are based on those tested maximums.

This paper assumes that the reader can set up and configure a Citrix Cloud environment according to recommended standards, with a minimum of three Cloud Connectors.

It is important to note that local host cache supports only on-premises StoreFront in each resource location or zone. In addition, local host cache supports server-hosted applications and desktops and assigned desktops. Local host cache is not supported for pooled desktops.

While outage mode is active, if the elected connector that brokers the sessions has an outage, the second connector becomes the elected high availability service. After the election, the second connector takes over to broker the sessions. The local host cache feature uses only one socket for multi-core CPUs for the connector VM configuration. In this scenario, we recommend a 4-core, 1-socket configuration.

Summary

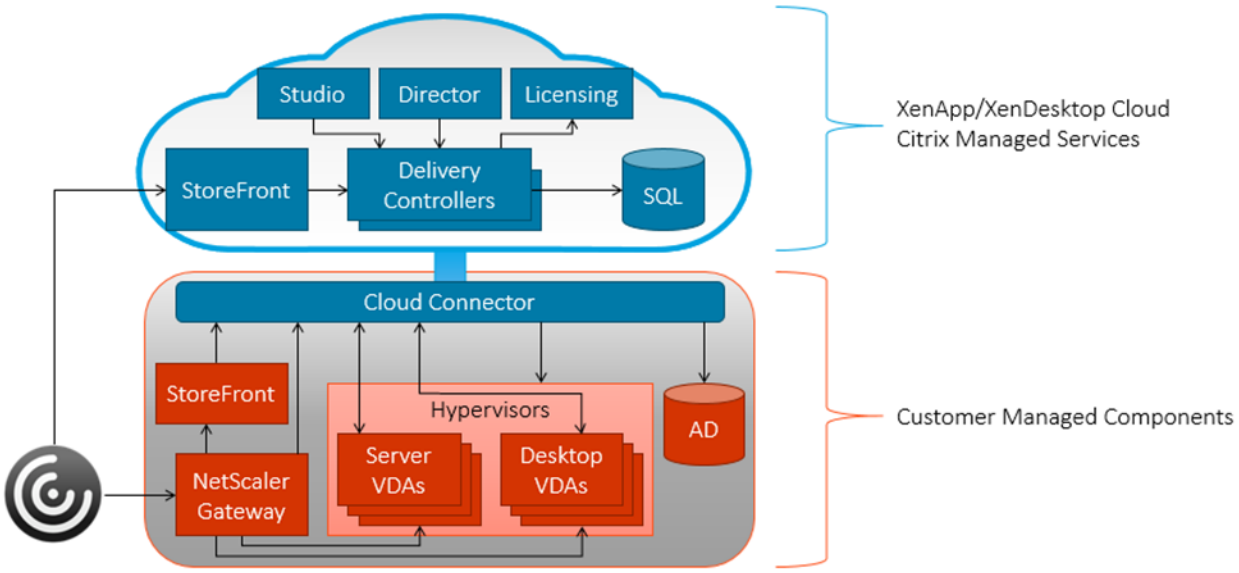
All results in this summary are based on the findings from test environments which we configured as detailed in the following sections. Different system configurations yield different results.

Key recommendations based on test results

- We recommend, for high availability sites that host no more than 5,000 workstations or 500 server VDAs, that you configure 3 VMs dedicated to the Cloud Connector. Each Cloud Connector VM requires 4 vCPU with 4 GB RAM. This configuration is an N+1 high availability configuration. Cloud connectors are deployed in high availability sets. Cloud connectors are not load-balanced. Because each CPU can process a limited number of connections, the CPU is the greatest limiting factor related to the number of workstations or server VDAs supported.
- Although this document focuses on testing with two Cloud Connectors, an N+1 set of three Cloud Connectors is recommended.
- We conducted session launch tests to compare local host cache outage mode active and inactive after a new configuration was synchronized and imported. The launch tests covered scenarios with 5,000, 20,000, and 1,000 session launches against the respective number of available workstations.
    - 5,000 sessions launched against 5,000 workstation VDAs
        - Tests used 2 Cloud Connector VMs, each had 4 vCPU and 4 GB RAM. Based on the recommendation for an N + 1 configuration, production environments should include 3 Cloud Connector VMs that meet these specifications.

        - Local host cache service peak consumed 91% of CPU resources and there was an average of 563 MB available memory
        - It took approximately 10 minutes from when the high availability service detected an outage for all VDAs to re-register with the high availability service, which is now the broker. We measured from the time the high availability service entered outage mode until the high availability service was ready to broker sessions again.
    - 20,000 sessions launched against 500 server VDAs
        - Tests used 2 Cloud Connector VMs, each had 4 vCPU and 4 GB RAM. Based on the recommendation for an N + 1 configuration, production environments should include 3 Cloud Connector VMs that meet these specifications.
        - Local host cache service peak consumed 90% of CPU resources and there was an average of 471 MB available memory
        - It took approximately 8 minutes from when the high availability service detected an outage for all VDAs to re-register with the high availability service. We measured from the time the high availability service entered outage mode until the high availability service was ready to broker sessions again.
    - 1,000 sessions launched against 1,000 workstation VDAs
        - Tests used 2 Cloud Connector VMs, each had 2 vCPU and 4 GB RAM. Based on the recommendation for an N + 1 configuration, production environments should include 3 Cloud Connector VMs that meet these specifications.
        - Local host cache service peak consumed 95% of CPU resources and there was an average of 589 MB available memory
        - It took approximately 7 minutes from when the high availability service detected an outage for all VDAs to re-register with the high availability service, which is now the broker. We measured from the time the high availability service entered outage mode until the high availability service was ready to broker sessions again.

Environment Overview



Citrix Cloud manages Cloud Connector services, and the customer manages the machines.

Test Methodology

We conducted tests by adding load and measuring the performance of the environment components:

- CPU
- memory
- database load
- Citrix Remote Broker Provider service
- Citrix High Availability service

We collected performance data, logon time, or both. In certain cases, proprietary Citrix simulation tools were used to simulate VDAs and sessions. The simulation tools are designed to exercise Citrix components the same way that traditional VDAs and sessions do, without the same resource requirements to host real sessions and VDAs.

Local host cache supports one elected high availability service per zone, not per site. For example, if you have five zones, one connector is elected as the broker in each zone. The Citrix Config Synchronizer service is responsible for importing the Citrix-managed site database. Every configuration sync creates a database, so initial configurations are needed, such as compiling stored procedures the first time the database is used. We executed all tests after a configuration sync.

Session Launch Tests

On customer-managed StoreFront servers, we started 5,000 and 20,000 session tests. The monitoring tools collect StoreFront log on time, resource enumeration, and ICA file retrieval.

Citrix uses simulation tools to facilitate high-volume user testing. The simulation tools, which are proprietary to Citrix, allow us to run the tests on less hardware than is required to run tests using real sessions at these levels (5,000 and 20,000 sessions). These simulated sessions go through the normal StoreFront log on, resource enumeration, and ICA file retrieval, but do not start active desktops. Instead, the simulation tool reports to the ICA stack that the session has launched and all communication between the broker agent and the broker service is consistent with that of an actual session. Performance metrics are gathered from Citrix Cloud Connectors. To determine how the environment responded to session launches, a sustained concurrency of 25 session launches was maintained at any given time throughout the duration of the test. The measurements therefore show results of a system under load throughout the test.

Test Results

Session Launch

The following tables compare session launch tests between local host cache outage mode active and local host cache outage mode inactive after a new configuration synchronization import. Each table shows the results for the number of sessions launched in the test.

5,000 workstation VDA sessions

	Local Host Cache outage mode Inactive (Normal Operations) Average Timing	Local Host Cache outage mode Active Average Timing
Authenticate	193 ms	95 ms
Enumerate	697 ms	75 ms
Total logon time	890 ms	170 ms
Retrieve ICA File	4,191 ms	156 ms

20,000 server VDA Sessions

	Local Host Cache outage mode Inactive (Normal Operations) Average Timing	Local Host Cache outage mode Active Average Timing
Authenticate	135 ms	112 ms
Enumerate	317 ms	91 ms
Total logon time	452 ms	203 ms
Retrieve ICA File	762 ms	174 ms

- 5,000 workstation VDA session launch test
    - There were approximately 30 ms of latency between the Citrix Cloud Connectors and Citrix Delivery Controller while local host cache outage mode was inactive.
    - There is a 720 ms difference in the logon process with local host cache outage mode active versus inactive, while the StoreFront is under load.
    - The largest time difference is in the retrieval of the ICA file, which is 4 seconds. This is largely because the connector is performing the brokering, whereas normally the StoreFront traffic traverses through the connectors to the Citrix delivery controller in Azure and back.
- 20,000 server VDA session launch test
    - There is a 249 ms difference in the logon process with local host cache outage mode active versus inactive, while the StoreFront is under load.
    - The difference in the retrieval of the ICA file is about 1 second.
- Compared to the 5,000-workstation VDA session launch, the 20,000-session launch test contains only 500 server VDAs, resulting in fewer calls from the Citrix delivery controller to the VDAs, which leads to lower response times.

Average CPU Usage Comparison

Session launch test		Average CPU %	Peak CPU %
5,000 workstation VDA sessions	Connector 1	8.3	38.2
	Connector 2	8.4	33.3
5,000 workstation VDA sessions - local host cache outage mode active	Connector 1 (elected high availability service)	42	**91**
	Connector 2	0.8	5
20,000 server VDA sessions	Connector 1	23	62
	Connector 2	23	55
20,000 server VDA sessions - local host cache outage mode active	Connector 1 (elected high availability service)	57	**90**
	Connector 2	0.8	6.6

- The table compares Citrix Cloud Connector CPU usage with local host cache outage mode active and local host cache mode inactive during 5,000 workstation VDA and 20,000 server VDA session launch tests.
- All Cloud Connectors are 4 vCPU and 4 GB RAM
- The elected high availability service machines peaked at 91% and 90% overall CPU respectively. It is worth noting that, while the non-elected high availability service does not have much usage, it may become the active if the elected high availability service has a failure. It is therefore critical for the connectors to have identical connector specifications.

Available Memory Usage

Session launch test		Average Available Memory (working set MB)	Peak Available Memory (working set MB)
5,000 workstation VDA sessions	Connector 1	636	657
	Connector 2	786	801
5,000 workstation VDA sessions - local host cache outage mode active	Connector 1 (elected high availability service)	563	618
	Connector 2	912	918
20,000 server VDA sessions	Connector 1	1030	1195
	Connector 2	1178	1329
20,000 server VDA Sessions - local host cache outage mode active	Connector 1 (elected high availability service)	471	687
	Connector 2	1210	1227

- The table compares available memory usage with local host cache outage mode active and local host cache mode inactive during 5,000 workstation VDA and 20,000 server VDA session launch tests.
- The number of sessions decreases the amount of available memory.
- There is a 54.35% (559 MB) increase in memory usage with 20,000 server VDA sessions when local host cache outage mode is active, mainly due to SQL server memory consumption.

Cloud Connector CPU Usage by Component

Session launch test	Component	Average CPU %	Peak CPU %
5,000 workstation VDA sessions	Connector 1 LSASS	2.4	10.7
	Connector 1 XaXdCloudProxy	3.5	18.5
	Connector 2 LSASS	2.5	12.9
	Connector 2 XaXdCloudProxy	3.5	21.2
5,000 workstation VDA sessions local host cache outage mode active	Connector 1 (elected high availability service) LSASS	12.9	29.5
	Connector 1 (elected high availability service) HighAvailabilityService	14.7	49.7
20,000 server VDA sessions	Connector 1 LSASS	7	12.2
	Connector 1 XaXdCloudProxy	8.7	15.5
	Connector 2 LSASS	7	12.5
	Connector 2 XaXdCloudProxy	9	15.7
20,000 sessions local host cache outage mode active	Connector 1 (elected high availability service) LSASS	4.3	17.2
	Connector 1 (elected high availability service) High Availability Service	4.5	18.2

- The preceding table shows the processes that consume the most overall CPU resources when local host cache outage mode is active, compared to when local host cache outage mode is inactive during 5,000 workstation VDA and 20,000 server VDA session launch tests.
- The Citrix Remote Broker Provider service (XaXdCloudProxy) is the top CPU consumer when local host cache outage mode is inactive.
- LSASS (Local Security Authority Subsystem Service) uses CPU during session logons. All authentications from Citrix-managed services must traverse the Citrix Cloud Connectors to communicate with the customer-managed Active Directory.

- The Citrix High Availability Service is used to broker the sessions, resulting in higher CPU usage when local host cache outage mode is active. Also, CPU usage peaked to 49.7% during the 5,000 workstation VDA session launch, while the usage was only 18.25% during the 20,000 server VDA session launch (500 VDAs). The difference is due to the number of VDAs.
- Connector 2 did not have any meaningful metrics, as it was not the elected high availability service.

VDA re-registration time while switching to local host cache

During a delivery controller outage, the 5,000 workstation VDAs must re-register with the elected local host cache broker. This re-registration time was ~10 minutes. The re-registration time for 500 server VDAs was ~8 minutes.

Number of VDAs	Re-Registration time
5,000 workstation VDAs	~10 minutes
500 server VDAs	~8 minutes

Outage Timings

Outage event	Number of VDAs	Time
Enter outage mode		10 minutes
Re-registration time to elected high availability service	500	~8 minutes
	5000	~10 minutes
Exit outage mode		10 minutes
Re-registration time to Citrix delivery controller	500	~5.5 minutes
	5000	~1.5 minutes

- There is a total of 20 minutes to enter (10 minutes) and exit (10 minutes) outage mode, due to the number of Citrix delivery controller health checks required. The time required to re-register the VDAs adds to the overall outage time.
- If the network is going up and down repeatedly, forcing an outage until the network issues resolve prevents continuous transition between normal and outage modes.

Database and high availability Service metrics with local host cache

Session launch test	Average high availability Service Database Transactions/sec	Peak high availability Service Database Transactions/sec
5,000 workstation VDA sessions	436	1344
20,000 server VDA sessions	590	2061

The preceding table shows the number of database transactions per second on the elected high availability service.

StoreFront CPU Usage Comparison

Session launch test	Average CPU %	Peak CPU %
5,000 workstation VDA sessions	4.5	32.4
5,000 server VDA sessions local host cache outage mode	13.8	32.6
20,000 server VDA sessions	11.4	22.1
20,000 server VDA sessions - local host cache outage mode	18.6	33.2

- The preceding table compares StoreFront CPU usage when local host cache outage mode is active to when local host cache mode is inactive during 5,000 workstation VDA and 20,000 server VDA session launch tests.
- The StoreFront machine has the following specifications: Windows 2012 R2, 8 vCPU (2 sockets, 4 cores each), 8 GB RAM
- When local host cache outage mode is active, there is approximately a 9% increase in average CPU usage with the 5,000 workstation VDA and about a 7% increase with the 20,000 server VDA session launch tests. The increase is mostly because the IIS worker processes more requests when local host cache outage mode is active. There is more CPU usage because StoreFront is processing session launches at a faster rate than when outage mode is inactive.

StoreFront Available Memory Usage Comparison

Session launch test	Average Available Memory (working set MB)	Peak Available Memory (working set MB)
5,000 workstation VDA sessions	5731	6821
5,000 workstation VDA sessions local host cache outage mode	5345	5420
20,000 server VDA sessions	4671	4924
20,000 server VDA sessions - local host cache outage mode	4730	5027

- The preceding table compares the StoreFront available memory usage when local host cache outage mode is active and when local host cache mode is inactive during 5,000 workstation VDA and 20,000 server VDA session launch tests.
- When local host cache mode is active, there is a 6.73% increase in memory usage during the 5,000 workstation VDA session launch test.

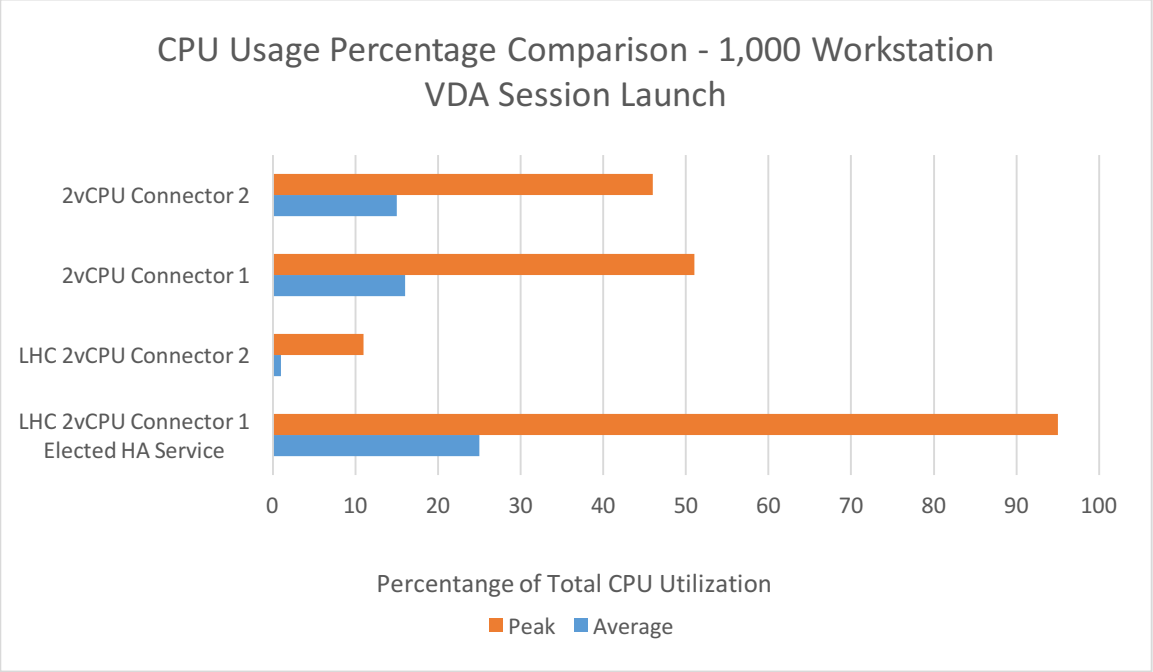
The following table compares outage mode active vs inactive after a new configuration synchronization import, launching 1,000 sessions to 1,000 workstation VDAs with local host cache, and using Citrix Cloud Connectors configured with 2 vCPU VMs.

Session Launch Comparison

	Local host cache outage mode inactive (normal operations)	**Local host cache outage mode active**
Authenticate	359 ms	89 ms
Enumerate	436 ms	180 ms
Total logon time	795 ms	269 ms
Retrieve ICA File	804 ms	549 ms

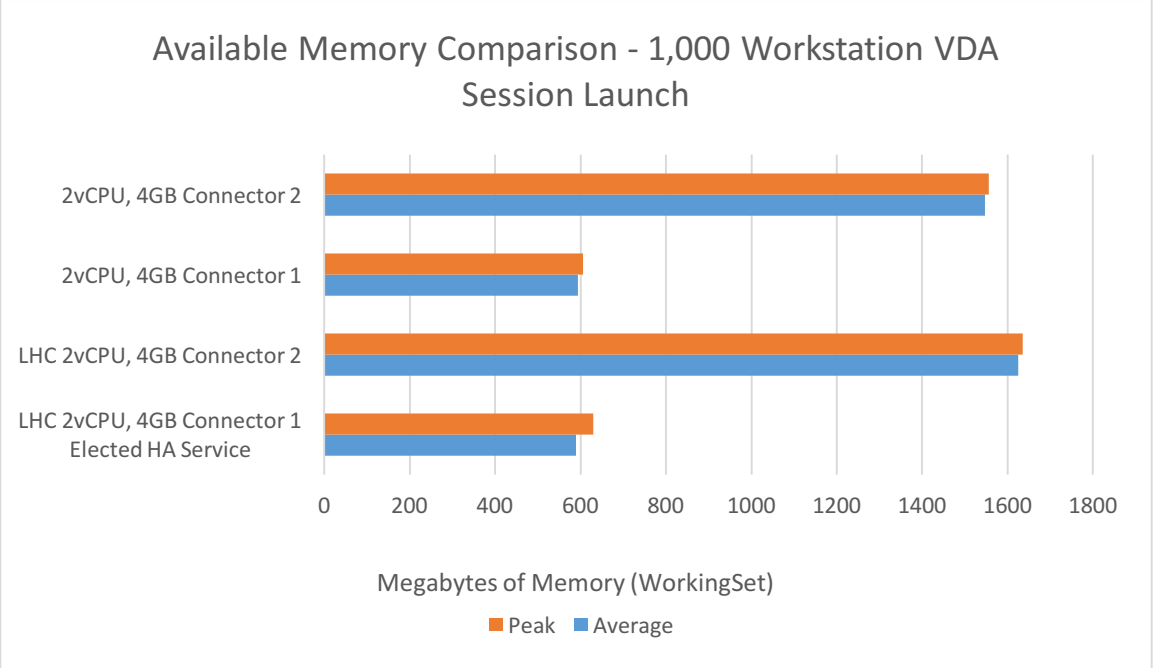
- While the StoreFront in under load, there is a 526 ms difference in the logon process when local host cache outage mode is active compared to when local host cache mode is inactive.
- There is a 255 ms difference in the retrieval of the ICA file when local host cache outage mode is active compared to when local host cache mode is inactive. The difference increases with the number of sessions.

Average CPU Usage Comparison



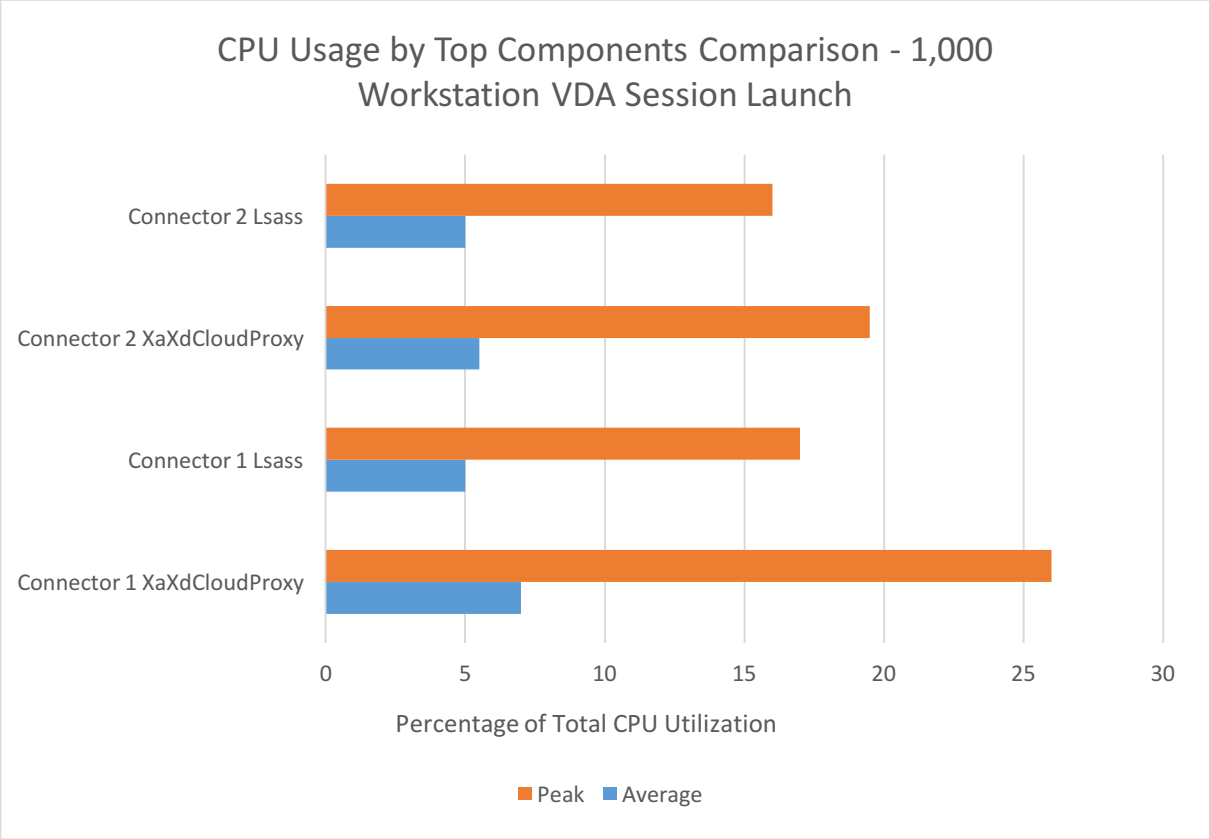
The elected high availability service peaked to 95% overall CPU, which indicates that 1,000 workstation VDA is an optimal configuration for a 2 vCPU connector VM.

Average Memory Usage Comparison

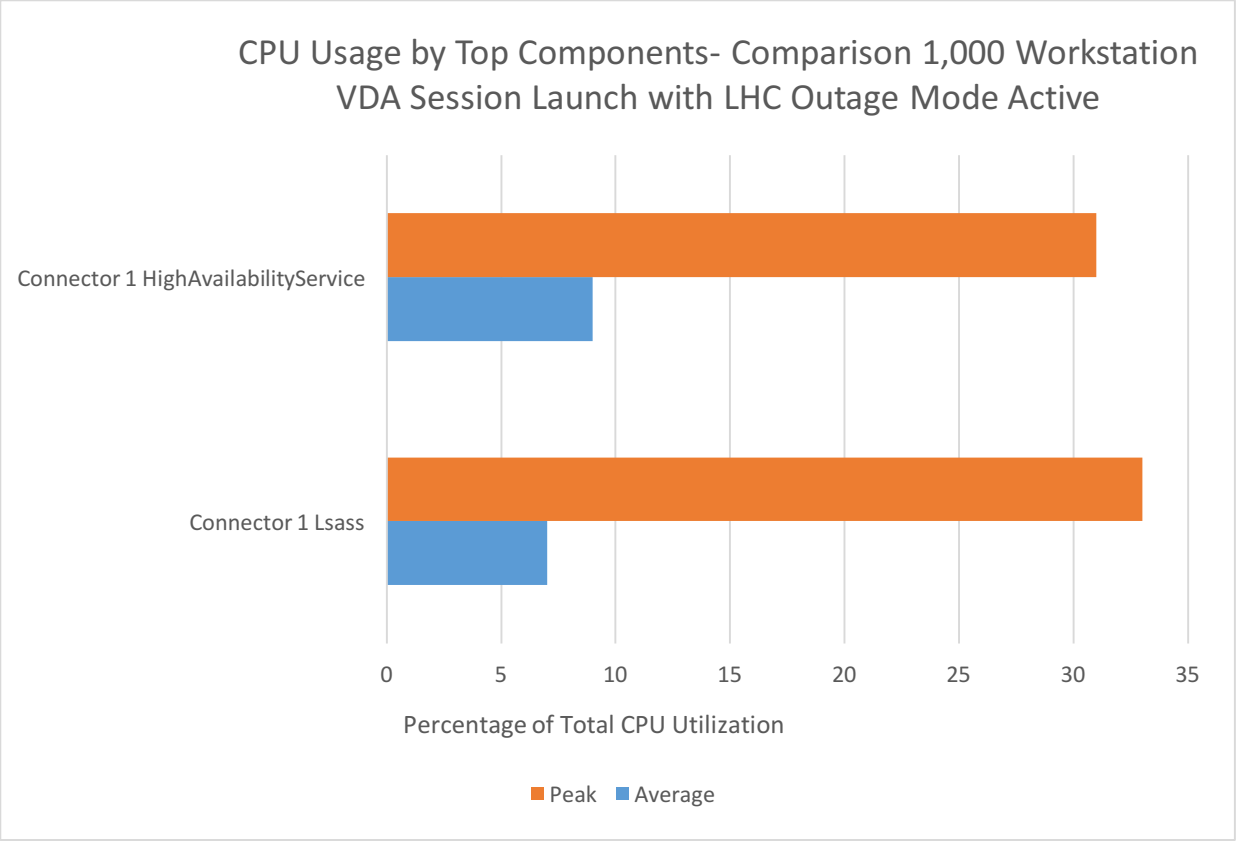


The preceding graph displays a comparison of Citrix Cloud Connector available usage when local host cache outage mode is active versus inactive, during a 1,000 workstation VDA session launch. There is not a significant difference in memory based on the local host cache outage mode.

Cloud Connector CPU Usage by Component Comparison



The preceding graph displays the processes that consume the most CPU resources while local host cache outage mode is inactive.

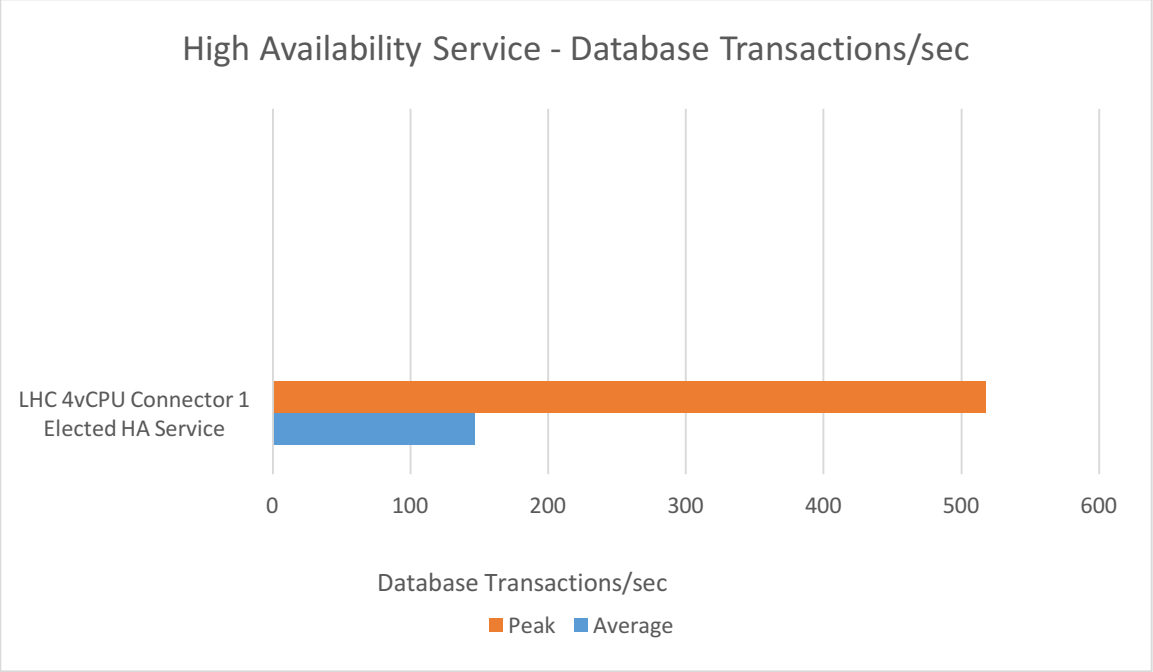


- The preceding graph displays the processes that consume the most CPU resources when local host cache outage mode is active.
- Connector 2 did not have any meaningful metrics.

VDA re-registration time while switching to local host cache

During a delivery controller outage, the 1000 workstation VDAs must re-register with the elected local host cache broker. The re-registration time was ~7 minutes.

Database and high availability service metrics with local host cache



The preceding graph displays the number of database transactions per second on the elected high availability service.

Impact with increasing number of zones on database import times

An extra zone (with a pair of its own connectors) was added to the test site to understand the impact. The first zone consists of 5,500 unique objects (2 catalogs). The secondary zone is a mirror of the first zone, and has its own unique objects, totaling 11,000 objects. It is important to note that local host cache is recommended only for zones with no more than 10,000 objects. Before we added the secondary zone, database import time on the connectors was about 4 minutes, 20 seconds. After we added the secondary zone and populated it with 11,000 objects, the import time increased to by ~30 seconds to ~4 minutes, 50 seconds. Adding more catalogs has marginal impact on import times. The largest contributing factors to performance degradation and increased import times are based on the number of assigned machines, users, and remote PCs. Additionally, 5,500 objects were split between 2 zones and the import time remained the same.

Number of zones	Total Number of Objects	Import time
1	5,500	4 minutes 20 seconds
2	11,000	4 minutes 50 seconds
2	5,500	4 minutes 20 seconds

Connector Sizing Guidance

For optimal performance, the following are the recommended configurations for Citrix Cloud Connector when local host cache mode is enabled.

Recommendation 1: to support 1,000 workstation VDAs using local host cache mode with Citrix Cloud Connector

- 2 Windows 2012 R2 VMs, each allocated with 2 vCPU (1 socket, 2 cores), 4 GB RAM
- This recommended sizing is based on the peak Citrix Cloud Connector overall 95% CPU usage and 589 MB average available memory while local host cache mode is active

Recommendation 2: to support 5,000 workstation VDAs OR 500 server VDAs using local host cache with Citrix Cloud Connector

- 2 Windows 2012 R2 VMs, each allocated with 4 vCPU (1 socket, 4 cores), 4 GB RAM
- This recommended sizing is based on
    - 5,000 workstation VDA sessions launched with local host cache mode active
        - Overall 91% peak CPU usage
        - 563 MB average available memory
    - 20,000 server VDA sessions launched with local host cache mode active
        - Overall 90% peak CPU usage
        - 471 MB average available memory

See the white paper Citrix Cloud XenApp and XenDesktop Service Sizing and Scalability Considerations for more information about general scalability sizing.

Test Environment

The test environment employed internally developed, proprietary testing tools, and VMs configured to the specifications in the following sections.

Tools Used

We used an internal testing tool to collect performance data and metrics from the machines under test and to drive the session launches. The in-house testing tool orchestrates user session launches to the XenApp and XenDesktop environment. The testing tool also provides a central location where we gather response time data and performance metrics. In essence, the test tool administers the tests and collects the results.

Test Configuration – XenApp and XenDesktop Service

The following is a list of the machine and OS specifications used with XenApp and XenDesktop Service testing.

Cloud Connectors:

- 2 Windows 2012 R2 VMs, each allocated 4 vCPU (1 socket, 4 cores), 4 GB RAM
- 2 Windows 2012 R2 VMs, each allocated 2 vCPU (1 socket, 2 cores), 4 GB RAM

StoreFront (Customer-Managed): Windows 2012 R2, 8 vCPU (2 sockets, 4 cores each), 8 GB RAM
Hypervisor: XenServer 7.0 + updates, 5x HP Blade BL 460C Gen 9, 2x Intel E5-2620 CPU, 256 GB RAM
Hypervisor Storage: 2 TB NFS share on NetApp 3250
VDA: Windows 2012 R2

Data Collection

We collect the following metrics from each test:

- average overall CPU, memory, component (cloud processes) usage increase
- VDA re-registration time when switching to the elected local host cache high availability service
- database and high availability service metrics when local host cache outage mode is active
- session launch comparison, average timings for
    - authentication
    - enumeration
    - ICA file retrieval
- impact to database synchronization times while increasing the number of zones
    - Time required to synchronize after a configuration change