A Sizing Study of Microsoft® Lync® Server 2010 on a Virtualized Dell™ PowerEdge™ R720

Make the most of Dell hardware running Microsoft Lync Server

Global Solutions Engineering

Dell

April 2012| Rev 1.0

Contents

Tables

Figures

Executive Summary

Microsoft Lync Server provides enterprise-grade communications for instant messaging, web/audio/video conferencing, application sharing, and telephony (or voice over IP). Users within an organization use the Lync client to connect to a Lync Server, and then use it to communicate with other users.

Virtualization is becoming increasingly important in many IT datacenters, and allows multiple operating systems or workloads to be installed on a single machine. By virtualizing the Lync Servers, IT administrators can:

- Take maximum advantage of available datacenter resources: with Intel® Xeon® E5-2600 product family processors offering up to 8 processing cores per CPU, and the latest Dell PowerEdge R720 supporting up to 768 GB of memory, having multiple Lync server components as separate VMs allows administrators to make use of hardware more effectively.
- Scale the infrastructure to run the Lync workload while minimizing the physical resources needed: virtualize multiple Lync Servers on a physical machine instead of restricting the IT datacenter's servers to a dedicated workload.
- Provide better availability: through the use of Microsoft's Hyper-V® failover-clustering, if one of the Lync Servers becomes unavailable, the Lync Server VM can be brought up, either running on the same physical machine or on another physical machine.

Keeping these advantages in mind, engineers at Dell's Global Solutions Engineering team conducted a scalability study of the Lync Server on Microsoft Hyper-V. The results show linear scaling when the number of heavy users was increased from 3000 to 6000 to 12,000 with 1, 2 and 4 VMs respectively on a single Dell PowerEdge R720.

Introduction, Scope, and Purpose

This paper begins with an overview of the Lync Server workload and the advantages of using Dell's latest R720 server, and then details the test environment and analyzes the collected performance metrics. Finally, based on the study's results, the paper presents a reference configuration for the Lync Server 2010 on PowerEdge R720 using virtualization.

This study benefits IT administrators and other professionals interested in using Microsoft Lync Server 2010 and Dell's 12th generation PowerEdge servers. This white paper analyzes the scalability of the PowerEdge R720 server with an increasing Lync user workload in a Hyper-V environment.

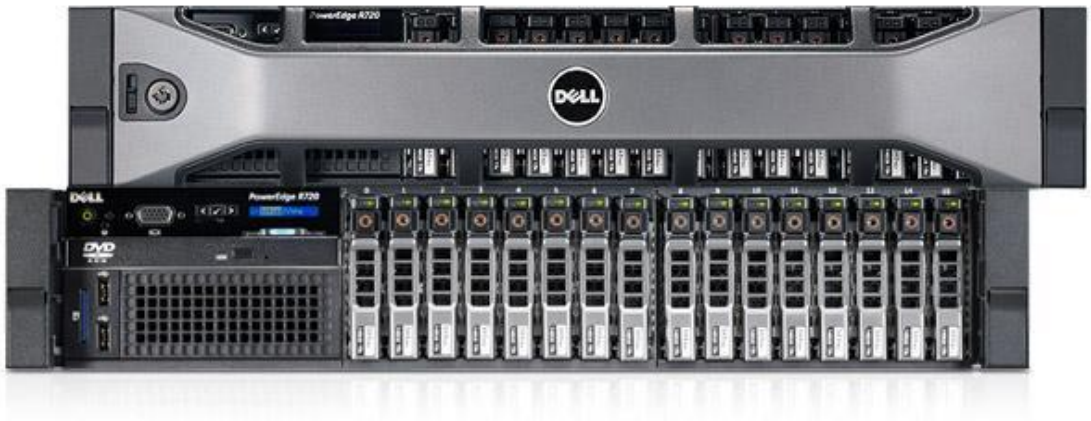
Advantages of Dell with Lync Server

The Dell PowerEdge R720 server features the latest Intel processors, highly scalable memory, and I/O optimizations that create a compelling building block for the Microsoft Lync Server 2010.

First, the R720 uses the new Intel Xeon E5-2600 processor product family. The processor's Intel Integrated I/O provides up to 80 PCIe lanes per server, and supports the PCIe 3.0 specification. In addition, a key feature included with the Intel Integrated I/O technology is the Intel Data Direct I/O (DDIO). Intel DDIO allows I/O traffic to skip the main memory and be directed straight to the processor cache. This redirection results in reduced latency and power consumption and increased bandwidth. Furthermore, the R720 has highly expandable memory: 24 memory slots with up to 32GB per DIMM, coming to a total memory capacity of 768 GB. The R720's flexible I/O capabilities allow it to handle the heavy I/O demands as well.

Lync supports audio/video, Web conferencing, instant messaging, VoIP, and other client features. These workloads can be computationally intensive, with audio/video traffic, Web conferencing, instant messaging, VoIP, and other client traffic moving simultaneously in an organization. The R720 flexible I/O capabilities allow it to handle the I/O demands of the Lync Server 2010. In fact, its reduced latency, improved bandwidth and reduced power consumption are critical for ensuring the quality of service (QoS) when a Lync deployment is scaled out.

Figure 1. Dell PowerEdge R720



For networking, Dell provides the Dell Force10 portfolio of top-of-rack, aggregation, core and distributed core switches. In the suggested reference configuration shown in Table 12 and Figure 9, two Dell Force10 S-Series 1Gbps top-of-rack switches are used – the Dell Force10 S55 and the Dell Force10 S60. Both these switches provide 1U top-of-rack 1/10 GbE connectivity, which is sufficient for the reference configuration. The S60 access switch provides deeper 1.25Gb buffer and is recommended for iSCSI SAN using Equallogic PS Series arrays in a redundant configuration. For the LAN, the low latency S55 access switch is recommended, also in redundant configuration. Both switches provide support for VLANs, ACLs, and management. They each contain 44 10/100/1000Base-T copper ports and 4 GbE ports that can be configured as copper or fibre.

Figure 2. Dell Force10 S55 and S60 TOR Switches

Storage recommendations leverage the Dell Equallogic PS6100 arrays in an iSCSI SAN. These arrays support 6Gb SAS bus speeds and have 4GB controller cache and four 1GbE ports (+ one Management port) per controller. Using PS6100XV, enterprises can leverage 15k SAS drives for their IOPS requirements. For lower IOPS requirements and larger capacity, enterprises can also consider PS6100X arrays with the 10K SAS drives.

Overview of the Lync Server

Microsoft Lync Server 2010 Enterprise Edition is a communications server solution that supports enterprise-level collaboration requirements. The Enterprise edition was selected for this study because it provides improved scalability and high availability. This scalable solution also includes a rich infrastructure that supports four different features for an integrated and unified user experience. These features are instant messaging (IM), application sharing, audio/video and web conferencing, and Enterprise Voice (VoIP).

In the Enterprise Edition of Lync Server, services that are installed together are consolidated. As a result, the number of server roles—a defined set of Lync Server 2010 functionality provided by a server—is reduced, resulting in reduced complexity during installation. Before Lync Server can be deployed, back end services such as Active Directory®, DNS and Microsoft SQL Server® must be functional. During deployment, a front end pool is created that consists of a set of front end servers—set of physical servers or set of virtual servers— that provide front end services. These services include Session Initiation Protocol (SIP) Registrar, SIP proxy, conferencing and other server services such as A/V conferencing, Web conferencing, instant messaging, application sharing, response group, bandwidth policy, call park, conferencing announcement and audio test service.

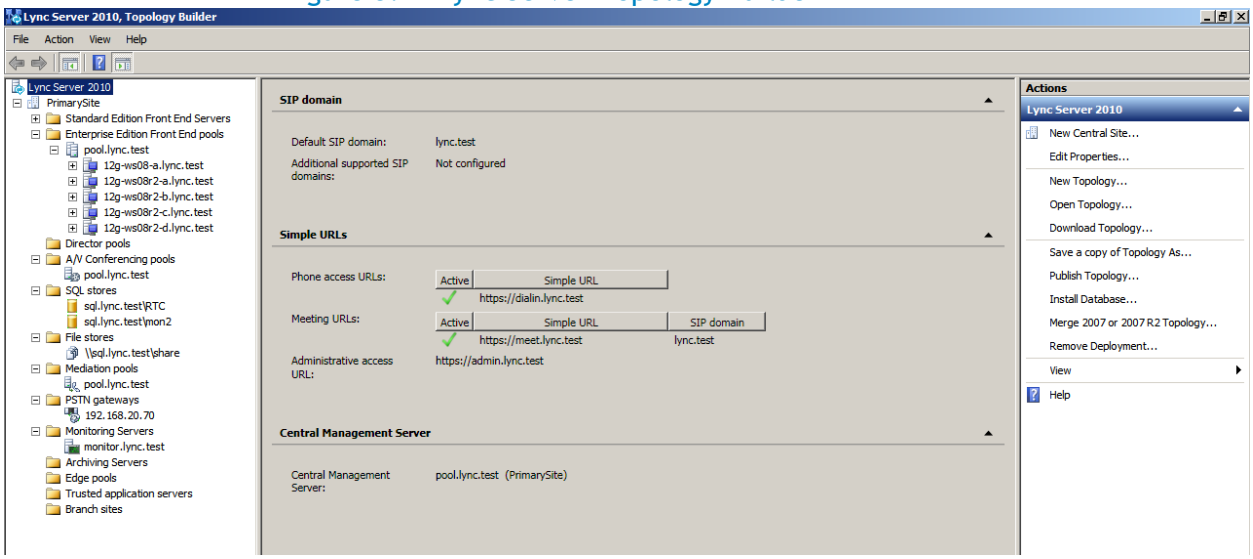
One advantage of a front end pool is the load balancing it performs on the front end servers; with load balancing, the number of client connections is evenly distributed across these servers. In the reference architecture described in this document, DNS load balancing is used for all the services and applications except Web traffic. For Web communication, a hardware load balancer is used instead. A load balancer is essential for high availability because it can redirect failed client connections, and also to ensure that each front end server in the front end pool is not overloaded.

The following subsections describe some of the Lync key features including the Topology Builder, Central Management Store, Lync Control Panel, and Lync Server Management Shell and the Lync Server roles. The back end services are also further described.

Topology Builder

The Lync Topology Builder manages the deployed Lync Server environment topology configuration. It can add components and roles to a temporary configuration file that is then published by saving the changes in a central database on the Central Management Store (CMS); the store is described in the next section. The server roles are installed by running the Lync Remote Setup Wizard on each server defined in the topology. The functionality of this wizard is not covered in this overview.

Figure 3. Lync Server Topology Builder



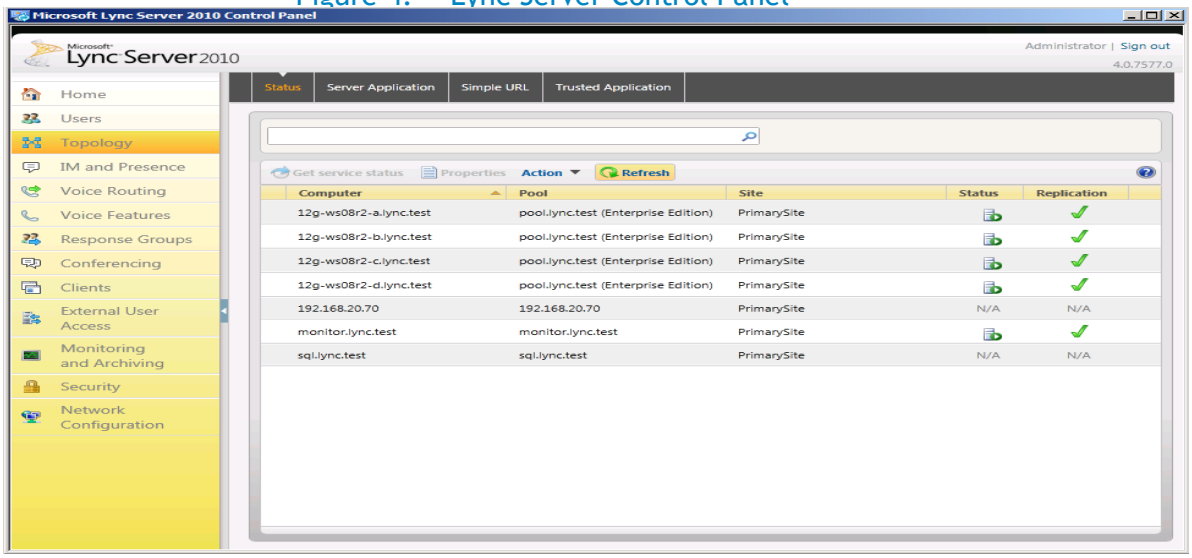
Central Management Store and Active Directory

Lync uses a new Central Management Store (CMS) that stores server and service configuration data. Individual user information, such as the user policy, the user's SIP URI, and the user's phone number, are stored in the CMS database. The CMS also provides data to the Lync Server Management Shell and file sharing. For backward compatibility with the deprecated Microsoft Office Communications Server 2007, Active Directory Domain Services (AD DS) contains only basic user information.

Lync Server Management Shell and Lync Control Panel

The Lync Server Management Shell contains 550 additional cmdlets to manage all aspects of a Lync Server deployment using PowerShell. In addition to this management shell, a graphical user interface (GUI) the Lync Control Panel, uses the Lync Server Management cmdlets as the underlying mechanism to perform management tasks, including the management of users in an organization. Figure 3 shows the Lync Server Control Panel. As seen in the figure along the left sidebar, it provides an interface for the management of Users, Topology, IM and Presence, Voice Routing, Voice Features, Response Groups, Conferencing, Clients, External Users, Monitoring and Archiving, Security, and Network Configuration. The Lync Control Panel replaces the Microsoft Management Console (MMC) snap-ins, the administrative interfaces of the older Microsoft Office Communications Server.

Figure 4. Lync Server Control Panel



Lync Server Roles

To test the scalability of Lync in a virtual environment, the Front End, A/V Conferencing, Mediation, Monitoring, and Back End server roles were installed. The Front End, A/V, and Mediation Server roles were collocated in the lab deployment, and these roles are described in the sections below.

Front End Server

The Front End Server role provides basic services for users. In the Enterprise Edition, a Front End Pool uses a group of servers that are configured identically and provide a similar set of resources; physical and virtual servers may not be mixed in a Front End Pool. The Front End Pool uses the Front End Servers in the pool to provide scalability and failover capabilities. Lync Server supports one or more Front End pools in a deployment, but only one pool can run the CMS.

The reference configuration, shown in Figure 8, uses a single Front End pool with four Front End Server Virtual Machines (VM). More details about the deployment of Lync can be found in the R720 System Configuration and Reference Configuration sections.

A/V Conferencing Role (Collocated with Front End)

Web conferencing enables users to view, share, and collaborate on documents, and to share their applications and desktops with each other. A/V conferencing enables users to communicate online with real-time audio. Either A/V and Web conferencing, or just Web conferencing, can be enabled when deploying conferencing. The reference configuration contains a recommendation that the A/V Conferencing role is collocated with the Front End Server role.

Some best practices call out the A/V conferencing role separately from the Front-End for configurations over 10,000 users. For the purpose of this study, and in order to analyze the scalability of Lync Server from 3,000 up to 12,000 users, the A/V Conferencing role was collocated with the Front-End.

Mediation Server (collocated with Front End)

This server role bridges Public Switched Telephone Network (PSTN) traffic to and from the media gateway to the Lync server network. It supports the routing of outbound calls to multiple media gateways, instead of a single media gateway as was the case in Office Communications Server. It also enables Media bypass that allows Lync clients and phones to directly route media traffic, excluding SIP traffic, to the media gateway without routing to the Meditation server. This role includes the Lync Server Meditation service and the Lync Server Replica Replicator Agent. This study recommends that the Mediation Server role be collocated with the Front End Server role in a similar manner as the Web and A/V Conferencing roles.

Monitoring Server

A monitoring server role can be deployed to collect statistical usage metrics for IM, conferencing, and Enterprise voice by tracking call detail records. It uses a back-end SQL database to store usage metrics through the SQL reporting services. For high performance, asynchronous messaging with Lync Server, the monitoring server depends on the Microsoft Windows® Message Queuing feature that guarantees message delivery, efficient routing, security and priority-based messaging. This feature must be installed on the monitoring server and Front End servers. Microsoft's best practices for Lync Server recommend that the Monitoring role be deployed on a separate server.

Back End Server

The Back End Server provides database services for the Front End pool. For most Lync deployments, a single database server is sufficient. In cases where failover is desired, then additional servers may be deployed to create a SQL Server cluster. It is recommended, as suggested in the reference configuration, to have multiple back end servers in a cluster.

Test Methodology

In order to determine the scalability of supporting multiple VMs on a single R720, a two-step approach was taken. First, the maximum number of users (using the Heavy profile in the Stress and Performance Tool) per VM was determined; it was found to be 3,000 users. Following these tests, additional VMs were added that had identical R720 host configurations until the solution could no longer scale due to the CPU, memory and other performance indicator thresholds.

R720 System Configuration

This study configured the R720 server using Microsoft and Dell best practices, taking into consideration Lync and hypervisor requirements. The R720 system was running Windows Server 2008 R2 with Service Pack 1, and the Hyper-V role was installed. The Lync Front End server role was installed on four VMs running Windows Server 2008 R2 with Service Pack 1.

To follow the established best practices, the Lync Front End VMs were SAN booted from the hypervisor onto a single LUN that resides on an EqualLogic SAN. These VMs were configured to use a non-legacy virtual network adapter and direct memory mapping. Because the Lync Server utilizes a large amount of network bandwidth, a total of nine 1 Gb Network links were used; four links were configured with multi-pathing I/O (MPIO) for SAN booting of the VMs, four ports were teamed for Lync network traffic, and two ports were used for Hyper-V management traffic. As an alternative, 3 ports can be used for

the LAN and the remaining single port can be used for management. This configuration will require only one additional add-on. The R720 host was installed with the Hyper-V role and no other roles in order to minimize the number of background processes.

The R720 memory and processors were critical in determining how well Lync scales. Because this study and reference configuration recommends the use of four front-end VMs, the R720 was provisioned with 96 GB of memory; each Lync virtual machine was allocated 4 vCPUs and 16 GB of statically assigned memory. Using 16 cores and 4 vCPUs per VM meant that a 1:1 ratio of total vCPUs to logical CPUs was maintained.

Table 1. Configuration of R720 (Lync VMs host)

Server	Dell PowerEdge R720
CPU	2 x Intel Xeon E5-2660 (8 cores @ 2.20 GHz)
Memory	96 GB
Operating System	Windows Server 2008 R2 SP1
VM Configuration	4 vCPUs and 16GB Memory

Tools Used for Testing and Validation

There were two main tools used for testing and validation: the Lync Stress and Performance Tool and the Windows Performance Monitor counters. The Lync Stress and Performance tool is written by Microsoft to generate a realistic load on a Lync system. The Windows Performance Counters provide fine-grained data for the Front End Server VMs and the Hyper-V host. Quality of Experience (QoE) reports from the Lync Monitoring role allow administrators to monitor good end-user call quality. Finally, performance counters from the load generation machines that run the Lync Stress and Performance tool validate that it is running correctly; these tools are explored in depth in the following sections.

Stress and Performance Tool

The primary tool used for sizing the Lync Server is the Lync Server Stress and Performance Tool from Microsoft. This tool can simulate the following end user features:

- Instant messaging: two-party communication between Lync clients using instant messages.
- Presence: updates to the user status (Available, Busy, Away, etc.)
- Audio, Application Sharing, and IM conferencing: conversations involving multiple parties using audio, instant messaging, and application sharing tools like Microsoft PowerPoint® or Excel®.
- VoIP calls using a PSTN simulator: VoIP calls can be made to and from the PSTN. For example, a call from a cell phone to a Lync user within the enterprise would be handled as an incoming PSTN call.
- Address book retrieval: one of the servers running the Lync Server in your deployment runs the ABS service. Lync clients download address books from the ABS to complete user look ups.
- Distribution List Expansion (DLX): Lync uses DLX to retrieve distribution list memberships that would consist of other Lync users.

It is important to note that the Stress and Performance Tool does not currently support video and Web conferencing. The Stress and Performance Tool was set up on multiple servers to generate the load on the Lync Server(s). The machines used for load generation were configured as follows:

Table 2. Configuration of Performance Tool Load Generators

Server	Dell PowerEdge R710
CPU	2 x Intel Xeon X5670 (4 cores @ 2.93 GHz)
Memory	72 GB
Operating System	Windows Server 2008 R2

The tests conducted on the R720 host that contained the Lync VMs were configured at the maximum load possible from the Stress and Performance Tool; the "Heavy" setting among 4 choices: "Heavy", "Medium", "Low", or "None."

Performance Counters from Front End Server VMs and Hyper-V Host

To collect more fine-grained data, performance counters were captured while running the Stress and Performance tool; these counters were collected on the Front-End and Host Hyper-V servers. Some of the important performance counters and thresholds used for the analysis are below.

Table 3. Host and Hyper-V Counters

Front End Servers (Hyper-V Virtual Machines)	
Performance Counter	Threshold
SIP Connections Active	>3000
Available Memory	>15%
Memory - Pages/sec	<500
Memory - Page Life Expectancy	>3600
AVMCU - Number of Conferences	Evenly Distributed across FE's
ASMCU - Number of Conferences	Evenly Distributed across FE's
DataMCU - Number of Conferences	Evenly Distributed across FE's
IMMCU - Number of Conferences	Evenly Distributed across FE's
DBStore - Queue Latency	<100ms
DBStore - SPROC Latency	<100ms
MCU Health State - AS	0 (Normal)
MCU Health State - AV	0
MCU Health State - Data	0
MCU Health State - IM	0
Average Holding Time for Incoming Messages	<10
Local 503 Responses/sec	~0
Local 504 Responses/sec	~0

Host Server (Windows Server 2008 R2 SP1 on PowerEdge R720)	
Performance Counter	**Threshold**
Network Utilization	<40%
Network – Output Queue Length	0
Available Memory	>15%
Processor Utilization (Logical Processor)	<60%
Processor Utilization (Hypervisor Root Virtual Processor)	<60%
Processor Utilization (Hypervisor Virtual Processor)	<60%
Disk sec/read	<15ms
Disk sec/write	<15ms

Initially, the tests were executed on a single Hyper-V virtual machine to establish the number of users that can be supported while maintaining performance thresholds. It was found that the Front End VMs running on the R720 could support 3000 users using the heavy profile for all the supported end-user features in the Stress and Performance tool. During the process, SIP connections to each Front End server were monitored to ensure that no connections were dropped as a result of bottlenecks in the server, storage, or networking. In addition, the metrics presented in Table 3 were all measured, and they verified that the system was within performance thresholds. These performance metrics are discussed below.

A value of 15% of available memory was used to identify issues related to a lack of memory. For memory pages, if a page has to be retrieved from the disk instead of from the memory, there is a negative impact to performance; the rate at which pages in memory are swapped with those in the disk needs to be below a 500 pages per second. If the rate is above this number, it indicates a lack of memory available to service requests quickly and will result in a substantial system slow-down. The page life expectancy can also indicate memory pressure, and anything below the threshold value of 3600 indicates insufficient memory.

To ensure that the tool was working and generating an acceptable load that is balanced across the entire Front End Pool, the number of conferences was recorded for Audio, Instant Messaging, Application Sharing, and Data Collaboration. To verify that none of the Front End Servers became overloaded during the tests, this counter was used in addition to the number of SIP connections.

The DBStore queue and sproc latency counters are essential for measuring bottlenecks within the back-end database; the queue counter represents the time taken by the backend database's queue for a particular request. The sproc counter represents the time taken for the backend database to actually process the request.

The MCU health counters give an indication of the overall system health; these should be 0 at all times, indicating normal operation. The average holding time for client transactions should be less than 3 seconds to allow up to 20 transactions per client; the 503 and 504 response counters should be close to zero. The 503 responses are more important because these counters indicate that the server is unavailable for client transactions, and 504 responses are more common and can be caused by an abrupt client logoff.

The primary indicators of the R720's performance are the processor, network, and memory utilization. Processor utilization can be impacted if measured from the Front-End VMs, because the CPU cycles are sliced for each VM, introducing latencies in the counters; for this reason the Hyper-V host's CPU utilization is used because it is not impacted. The Logical Processor counters give the total CPU usage running on all available machine cores. The hypervisor root virtual processor counters measure the CPU utilization for the Hyper-V host OS, and the hypervisor virtual processor counters measure the CPU utilization for the VMs. These counters suffer a slight amount of clock impact, and can exceed 100%. For networking, we made certain that there was sufficient bandwidth across the teamed NIC and no queues were impacted due to network congestion. Memory utilization was monitored to ensure that there were no I/O bottlenecks.

Quality of Experience Reports

Quality of Experience (QoE) is an important parameter for any real-time communication, and the Lync Server provides a Monitoring Server role that can analyze the QoE metrics of calls taking place over a fixed time period. For this study, a time period of 8 hours was selected for analysis, and the QoE indicators measured are in the following table.

Table 4. Monitoring Server QoE Statistics and Thresholds

QoE Metric	Threshold
Jitter	< 20ms
Packet Loss	< 0.1%
MOS	< 0.5
Round Trip Time	< 200ms

Across TCP/IP networks, packets can arrive from one Lync client to another at irregular intervals, causing jitter, and packets can also be lost in the network leading to poor call quality. The MOS metric measures the degradation of VoIP calls in the Lync system using computer algorithms. The round-trip time is the time it takes for a packet to travel from one client to another in addition to the receiver's acknowledgement to the transmitter; high round-trip times indicate choppy voice quality.

Load Generation Performance Counters

In addition to the counters from the Lync VMs and their host machine, counters from the load generators were also collected to verify that the load generation system did not introduce latencies. The major performance counters are in the Load Generation Counters Table below.

Table 5. Load Generation Counters

Performance Counter	Threshold
CPU Utilization	<60%
Available Memory	>15%
Network utilization	<50%

On the client machines, the CPU, memory, and network utilization were set below the acceptable limits of the Hyper-V Host, so that the desired load can be generated for the Lync Servers. To verify that the Lync system was healthy, the following client counters were also recorded:

- Total Active Endpoints
- Presence Pass Rate Percentage
- Total IM Calls Active
- Total Number of IM Conferences Active
- Total Number of AV Conferences Active
- Total Number of AS Conferences Active
- Total Number of Data Conferences Active
- Distribution List Calls per second
- CAA Calls in progress

Results and Analysis

Three test scenarios were run, and the results were collected from the Front-End VMs, Host, Monitoring Server Reports and the Stress and Performance Tool counters. The three major scenarios in the testing included:

Table 6. Test Scenarios

Scenario	Number of VMs	Total Number of Users	Users per VM	Hostnames of running VMs
Scenario A	1	3000	3000	FrontA
Scenario AB	2	6000	3000	FrontA, FrontB
Scenario ABCD	4	12000	3000	FrontA, FrontB, FrontC, FrontD

In Scenario A, one front end VM – named FrontA – was running with a total of 3000 heavy users. In Scenario AB, FrontA from Test A was running with a 3000 users, and an additional VM named FrontB was added that ran services for an additional 3000 users, creating a total of 6000 heavy users. In Scenario ABCD, two more virtual machines were added, each supporting 3000 additional users for a total of 12,000 heavy users.

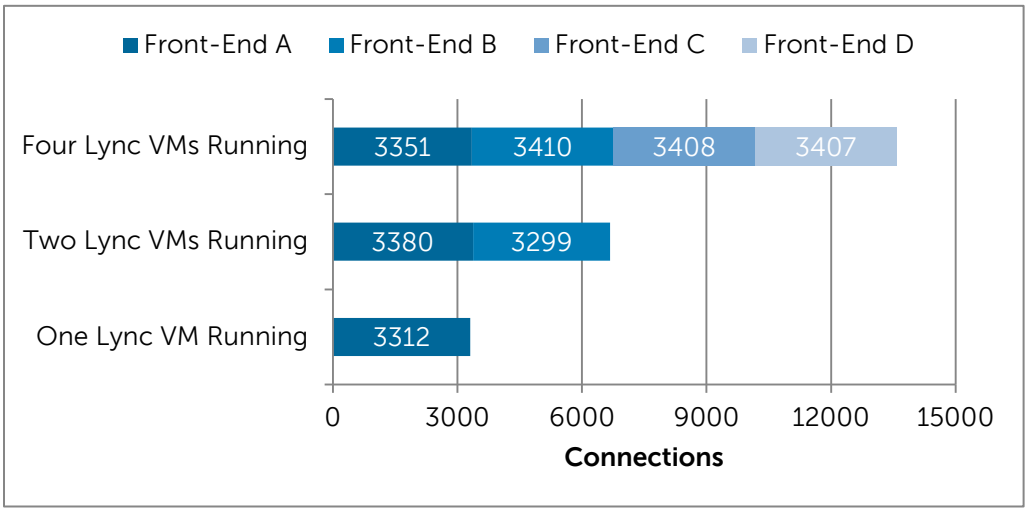
The Reference Configuration uses four Front End servers across two R720 Hyper-V hosts. If one of the R720 Hyper-V hosts goes offline, the two Front End servers running on it can be migrated to the other operation R720 Hyper-V host. In that instance, there would be a worst case of four Front End VMs running on a single Hyper-V Host. The three test scenarios were conducted on a single R720 Hyper-V host. Scenario AB represents normal operation in the Reference Configuration (Figure 9) and Scenario ABCD represents the worst case scenario in the Reference Configuration. To study the scalability of the R720 platform with increasing Lync Load, “Scenario A” investigated the performance of a single VM.

To explain our analysis of these results, first the Front End Server VMs performance is discussed, followed by the results of the Hyper-V host's performance. Third, the Lync QoE thresholds are verified to be within acceptable levels, and finally, the counters from the Stress and Performance Tool are also verified as within acceptable levels.

Results from the Front End Server VM(s) Counters

To show that the Front End Servers are performing correctly, the Lync load needs to be balanced across all running Front End Servers and then the performance counters verified to be within acceptable levels. To ensure that 3000 users were connected to each Front End VM, the user connections to the Front-End servers were logged; the results from the counters indicate that at least 3000 users were logged in, which can be seen in the Lync User Connections per VM chart below.

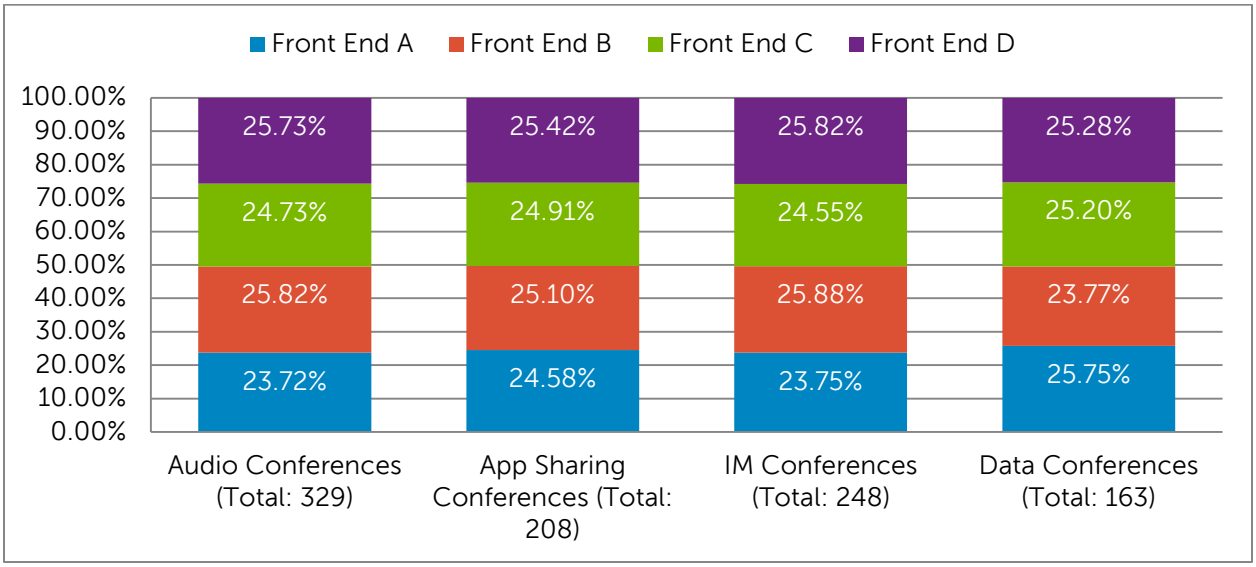
Figure 5. Lync User Connections per VM



The "Four Lync VMs Running" bar shows a near even distribution of around 3000 users connected to each Front End VM. The "Two Lync VMs Running" bar also shows an even distribution. The "One Lync VM Running" bar verifies that there are approximately 3000 users. The number of user connections is greater than 3000 because users connect and disconnect to different Front End Servers during the test; the connection balance demonstrates that DNS Load Balancing worked effectively in distributing the clients amongst the VMs.

In addition to client connections, conferences should be evenly distributed across the Front End VMs. For 12,000 users, the total number of conferences and their distribution among the four front-ends are shown below. There were a total of 329 audio conferences, 208 app-sharing conferences, 248 IM conferences, and 163 data conferences all running concurrently.

Figure 6. Conference Distribution per Front-End for 12,000 users



As can be seen above, the distribution of conferences is evenly distributed between the four VMs. The table below summarizes the results from the Front End VMs counters.

Table 7. Lync Counter Thresholds for Front End VMs

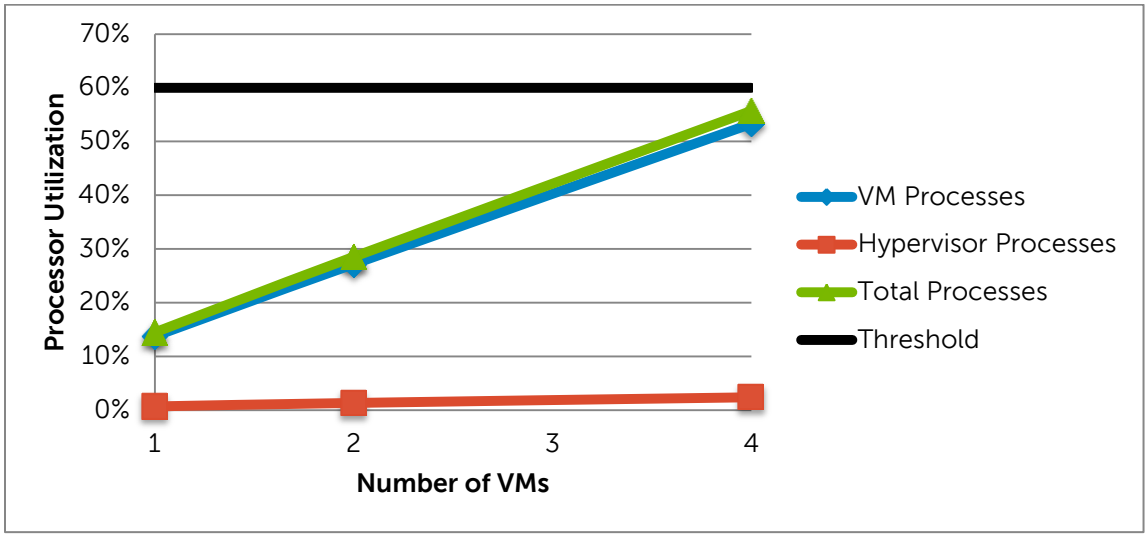
Performance Counter	Threshold	1 VM	2 VMs	4 VMs
LS:SIP - 01 - Peers(_Total)\SIP - 000 - Connections Active	>3000	3312	3339	3393
Available Memory	>15%	61%	63%	62%
Memory\Pages/sec	<150	0.16	1.24	0.36
Page Life Expectancy	>3600	16830	17214	18538
SIP - Local 503 Responses/sec	~0	0	0	0.01
SIP - Local 504 Responses/sec	~0	0	0	0
SIP - Average Holding Time For Incoming Messages	<10	0.27	0.11	0.46
DBStore - Queue Latency (msec)	<100	1.26	1.42	18.75
DBStore - Sproc Latency (msec)	<100	5.93	7.71	16.9
MCU Health State (AS, AV, Data, IM)	0	0	0	0
SIP - Average Holding Time For Incoming Messages	<10	0.27	0.11	0.46

All of these numbers were taken from the eight hour tests, ignoring the initial period during which clients log-in to the Front End(s). All metrics are within the necessary thresholds, showing that the Lync Server deployment performed correctly.

Results from Host Counters

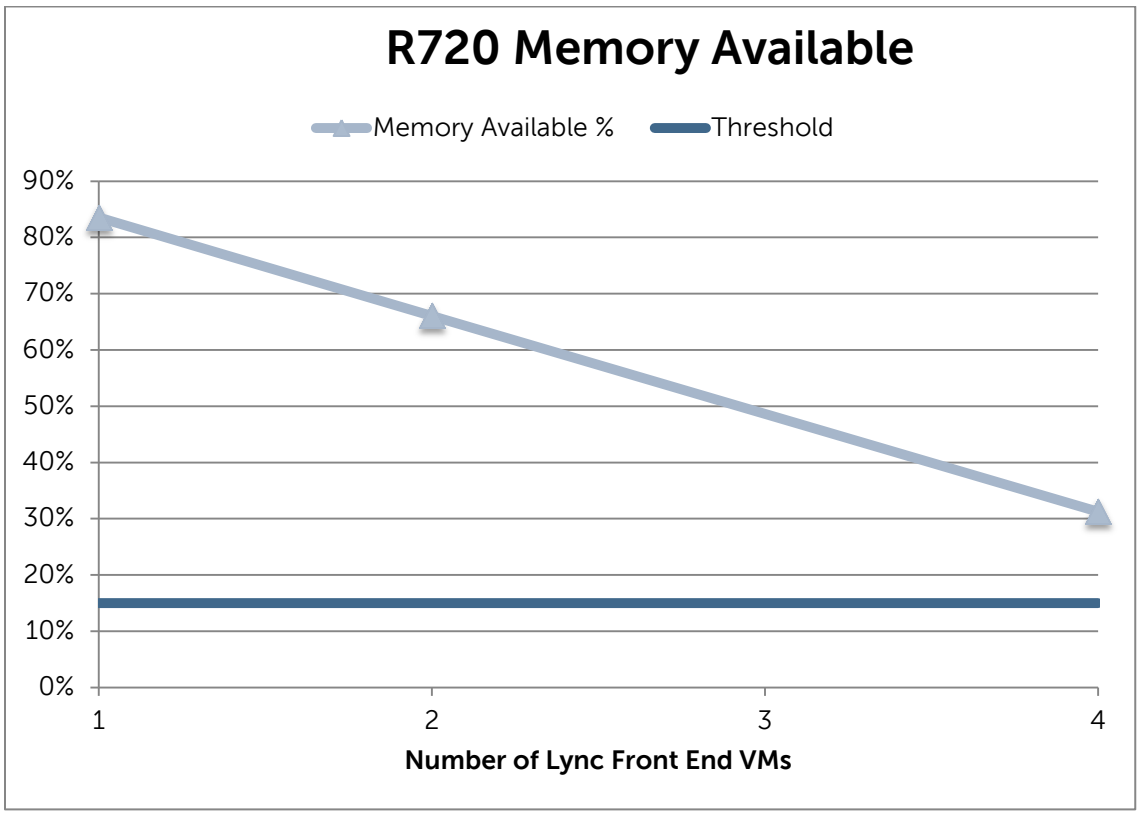
As mentioned previously, the processor counters are best measured from the Front End VMs host. These measurements are shown in the chart below.

Figure 7. Processor Utilization and Scaling as Lync Users Increase



In the chart above, there is a linear scaling of CPU usage when the user load is increased from 3000 users on one VM to 6000 users on two VMs, and then to 12,000 users on four VMs. At 12,000 users, the observed CPU usage over the eight hour test period was 56%. The chart below shows the available memory when the load is increased.

Figure 8. Memory Availability and Scaling as Lync Users Increase



Again, the available memory shows a linear relationship to the number of users supported by the VM's. As expected, the greater the number of users (and VM's) on the same Hyper-V host decreases the available memory; the available memory is, however, not close to the threshold set earlier of 15%.

At 12,000 users, the teamed NIC proved to be sufficient. The final parameter measured was the disk latency for the virtual machines residing on the EqualLogic® storage unit. VM storage latency can have an impact on the machine's performance, as is shown below and was found to be within acceptable limits.

Table 8. Storage Latency for Hyper-V VMs on SAN (12,000 users)

Latency Counter	Threshold	Measured
Disk sec/read	<15ms	8.89ms
Disk sec/write	<15ms	5.62ms

Quality of Experience Results

The tables below summarize the QoE results captured from the monitoring server, which indicates that the deployment is in a healthy state.

Table 9. QoE Summary for Peer-to-Peer Calls

QoE Metric	Threshold	3000 users	6000 users	12,000 users
Jitter	< 20ms	0.17ms	1ms	1ms
Packet Loss	< 0.1%	0	0	0
MOS	< 0.5	0.02	0.03	0.04
Round Trip Time	< 200ms	0	0	0.02

Table 10. QoE Summary for Conferencing

QoE Metric	Threshold	3000 users	6000 users	12,000 users
Jitter	< 20ms	1ms	1ms	1ms
Packet Loss	< 0.1%	0	0	0
MOS	< 0.5	0.07	0.08	0.08
Round Trip Time	< 200ms	1ms	1ms	1ms

As shown above, both peer-to-peer and conference scenario statistics are within acceptable QoE limits.

Stress and Performance Tool Counters

To ensure that the Stress and Performance tool running on the load generator servers did not experience bottlenecks, the following counters were measured.

Table 11. Stress and Performance Tool Counters

Performance Counter	3000 users	6000 users	12,000 users[1]
Processor Utilization	< 20%	< 60%	< 60%
Available Memory	> 50%	> 50%	> 50%
Network Utilization	<1%	<1%	<1%
Total Active Endpoints	3236	6492	6500
Presence Pass Rate %	100	100	100
Total IM Calls Active	790	1584	1588
Total IM Conferences Active	44	89	88
Total AV Conferences Active	34	69	68
Total AS Conferences Active	21	42	43
Total Data Conferences Active	7	7	14
Total DLX Calls / Second	0	1	1
CAA Calls in Progress	8	18	9

As can be seen from the table above, the processor, memory, and network utilization were well within the thresholds and the resources were adequate, indicating that the load was generated on the Front End VMs without any bottlenecks.

Reference Configuration

Based on the tests, a suitable highly-available reference configuration is outlined below. The configuration takes into account the additional overhead of one of the Front-End VM's or hosts experiencing a failure, or needing to be brought down for maintenance. The back-end SQL database is configured in a two-node SQL cluster; there is an additional server allocated for Archiving/Monitoring. The table and diagram following summarize the suggested configuration.

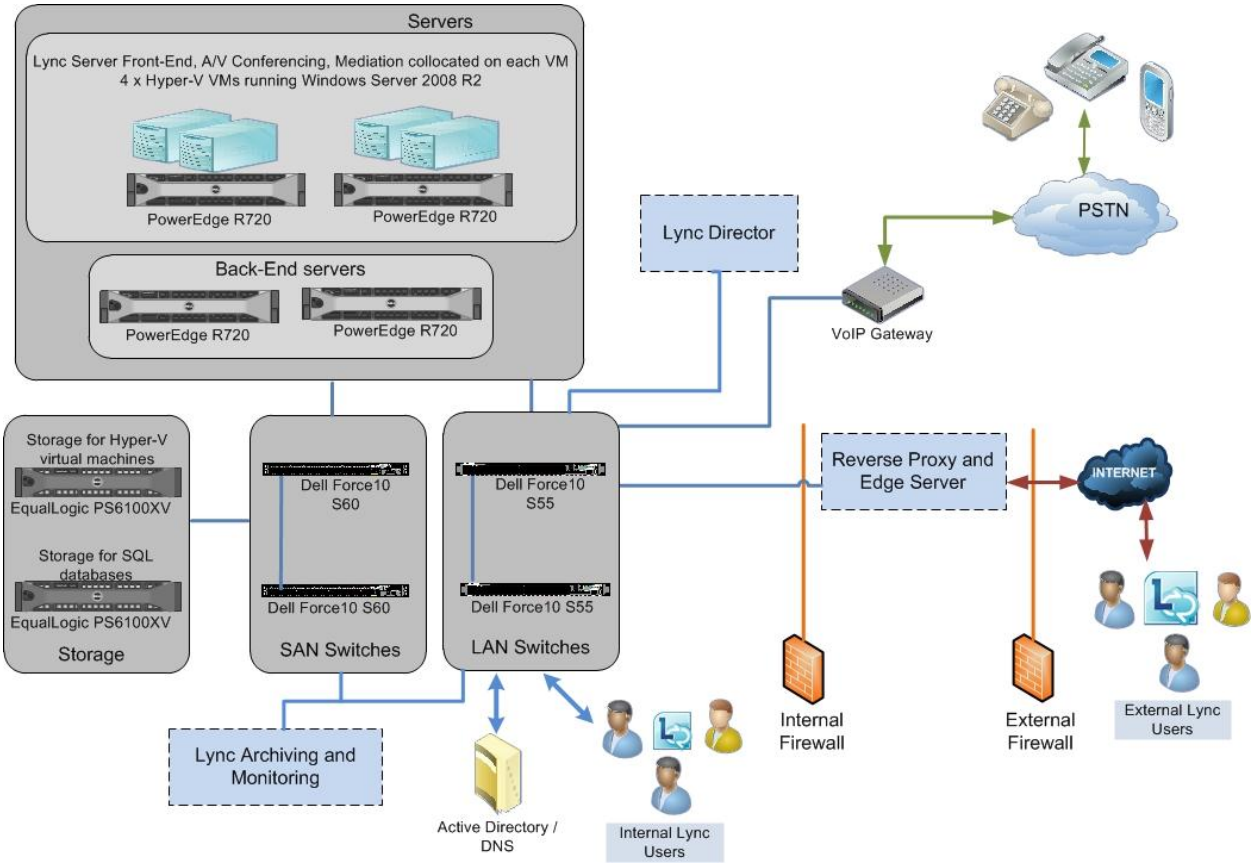
[1] An extra load generator was used for the 12,000 users scenario. For total numbers and CAA calls in progress, multiply by 2.

Table 12. Reference Configuration for 12,000 Users

Server Configurations	Detail
Microsoft Lync Server Version	Enterprise Edition
Physical Server Configuration (Host)	2 x R720 2 x 8-core Intel Xeon 24 x 4 GB = 96 GB Memory 2 x 146GB 15k SAS
VM Configuration: Front End, Mediation and A/V Conferencing Server Roles	4 x Hyper-V Windows Server 2008 R2 VM 2 x VMs per host 4 vCPUs per VM 16 GB Memory per VM
Back-End Server	2 x R720 (in a fail-over cluster) 2 x 6-core Intel Xeon 16 x 2 GB Memory = 32 GB
Storage Configuration	**Detail**
Storage for Hyper-V VM's	Dell EqualLogic PS 6100XV iSCSI SAN 24 x 146GB 2.5" NL-SAS in RAID 10
Storage for Back-End Database, Archiving/Monitoring Database	Dell EqualLogic PS 6100XV iSCSI SAN 24 x 146GB 2.5" NL-SAS in RAID 10
Additional Hardware	5 x Quad Port Network Interface Cards[2]
Networking Configuration	**Detail**
LAN Networking	2 x Dell Force10 S55 Switches
SAN Networking	2 x Dell Force10 S60 Switches
VoIP Connectivity	PSTN Gateway or SIP Trunking
Optional Components	**Detail**
Additional Server Roles	Lync Server Director Pool Lync Server Archiving and Monitoring

[2] Connectivity to EqualLogic iSCSI SAN for 1 x Archiving/Monitoring, 2 x Back-End, 2 x Front-End Hosts

Figure 9. Reference Architecture for 12,000 users on Dell PowerEdge R720



The storage and networking are enabled by Dell EqualLogic PS6100 series arrays, with 15k SAS drives and Dell Force10 S60/S55 switches respectively.

Conclusion

This paper presented testing results from a virtualized Microsoft Lync Server 2010 deployment on the Dell PowerEdge R720 server. It discussed the new capabilities on the R720, and how the Lync Server scales well with these new hardware features. By collecting data from the Lync Stress and Performance Tool and performance counters, the almost linear scaling of Lync Server 2010 was observed. Care was taken to validate that the Lync deployment and load generation tool were within good performance thresholds, and a Reference Configuration was created based on the Lync server's performance.

The Reference Configuration uses a SAN, physical backend servers, and two Hyper-V hosts to run four Lync Front End Server VMs. It uses the Dell PowerEdge R720 server, Force10 S60 and S55 switches, and EqualLogic PS6100XV storage. The test results show that even in the worst case scenario, Lync will perform within performance thresholds using the R720 and the resource utilization scales linearly with increasing load on the virtual machines.