Scaling Exchange 2013 Workloads on Dell EqualLogic PS6210E and Windows Server 2012 R2 with Hyper-V

Dell Engineering
July 2014

A Dell Reference Architecture

Revisions

Date	Description
July 2014	Initial release

Table of contents

Acknowledgements

Feedback

We encourage readers of this publication to provide feedback on the quality and usefulness of this information by sending an email to SISfeedback@Dell.com.

SISfeedback@Dell.com

Executive summary

Handling unpredictable factors while designing and planning an infrastructure that supports a messaging solution is a constant threat to the subsequent day-by-day operations. An increasing in user demand, preparing for growth and organization mergers or acquisitions can suddenly and exponentially raise the computing and storage demand from the original plan.

Dell EqualLogic provides an easily scaled-out storage platform on a versatile virtualization infrastructure supported by Microsoft Windows Hyper-V technologies. This solution is able to address such challenges through a modular or building block approach.

The ability of an EqualLogic storage solution to seamlessly scale from one to multiple arrays provides an easy-to-deploy resolution for a sudden or planned change in the storage demand of both capacity and performance. Virtualization technologies assist with growth by providing automation mechanisms and flexibility in provisioning new resources and in distributing the workloads across the virtual infrastructure.

The solution presented in this paper focuses on different aspects of scaling a messaging workload of Microsoft Exchange Server 2013 in a virtualized infrastructure based on Hyper-V technologies and with a back-end storage on EqualLogic PS6210E arrays with 4 TB drives.

- A comparison between a host and a guest iSCSI initiator performance for the virtualized workload
- The differential of performance when scaling up the user count on a storage unit with either the guest or host initiator
- The storage response progression when scaling out the solution due to an increase in user count, expanding the SAN to multiple arrays

1 Introduction

The reference architecture validated for the purpose of this paper selects some of the most common planning variables (mailbox size and user profile) for a Microsoft Exchange solution. The deployment prepared for an increasing number of users and enabled scaling the workload up and out on the back-end SAN. Besides these exercises, an assessment was made of the variance in response to the iSCSI initiator while deployed from a guest or directly from the host server.

1.1 Purpose and scope

This paper is primarily intended for IT professionals (IT managers, Solution Architects, Exchange and Storage Administrators, and System and Virtualization Engineers) who are involved in defining, planning, or implementing Exchange Server infrastructures and who would like to investigate the benefits of using EqualLogic storage. This document assumes the reader is familiar with Exchange Server functions, EqualLogic SAN operation, and Microsoft Hyper-V architecture and system administration. The scope of this paper is restricted to a local datacenter topology and does not include specific or detailed server sizing information.

1.2 Terminology

The following terms will be used throughout this document.

Group: Consists of one or more EqualLogic PS Series arrays connected to an IP network that work together to provide SAN resources to host servers.

Member: Identifies a single physical EqualLogic array.

Pool: A logical collection that each member (array) is assigned to after being added to a group that contributes its storage space to the pool.

Hypervisor: Denotes the software layer that manages the access to the hardware resources, residing above the hardware, and in between the operating systems, running as guests.

Virtual Machine (VM): An operating system implemented on a software representation of hardware resources (processor, memory, storage, network, and others). VMs are usually identified as guests in relation with the host operating system that executes the processes to allow them to run over an abstraction layer of the hardware.

Balanced tree (B-Tree): A tree data structure where a node can have a variable number of child nodes; commonly used in databases to maintain data sorted in a hierarchical arrangement. It allows efficient data access to the pages for insertion, deletion, and searches.

Key performance indicators (KPI): A set of quantifiable measures or criteria used to define the level of success of a particular activity.

2 Dell EqualLogic PS6210 arrays for Exchange Server 2013 deployments

Simplified, unified, virtualized EqualLogic storage is ideal for companies and organizations with growing data and performance needs. The EqualLogic frameless, IP-based peer storage architecture enables live, non-disruptive firmware upgrades and maintenance, as well as linear scaling of capacity and performance with each array added to the group.

Dell new generation EqualLogic PS6210 Series arrays provide enhanced storage performance for the small-to-medium enterprise with the simplicity of the EqualLogic product line. Powered by the updated 7.0 firmware, the PS6210 family of high-performance arrays is ideal for virtualization, databases, consolidation/collaboration and heavy I/O applications. These virtualized iSCSI SANs combine intelligence and automation with fault tolerance to provide simplified administration, rapid deployment, enterprise performance and reliability, and seamless scalability using innovative Fluid Data technology.

The EqualLogic PS6210E with multiple high capacity drive size options, up to 4 TB 7.2K RPM drive capacity with a Self-Encrypting Drive (SED) option, provides both entry-level configuration for affordable capacity and high capacity configurations for data-intensive applications, highly consolidated environments and collaboration applications such Microsoft Exchange 2013. The PS6210E with 4 TB drives offers IT generalists the ability to manage more data with fewer resources, capability to integrate tightly with Exchange 2013 environments and the flexibility to use with various operating systems/hypervisors like Microsoft Hyper-V Server 2012 R2. The PS6210E offers low TCO that can be tailored to fit exact messaging needs, and grow simply and cost effectively.

3 Solution architecture overview

The solution presented and evaluated in this paper is built on a virtual infrastructure supported by Microsoft Windows 2012 R2 with Hyper-V and a back-end iSCSI SAN provisioned by Dell EqualLogic Storage. The operating system of the VMs simulating the Microsoft Exchange workload and of the remaining VMs running the infrastructure and monitoring is Windows Server 2012 R2.

3.1 Conceptual system design

The elements of the infrastructure supporting the simulated environment, their main relationships and connectivity links are represented in the conceptual diagram in Figure 1.
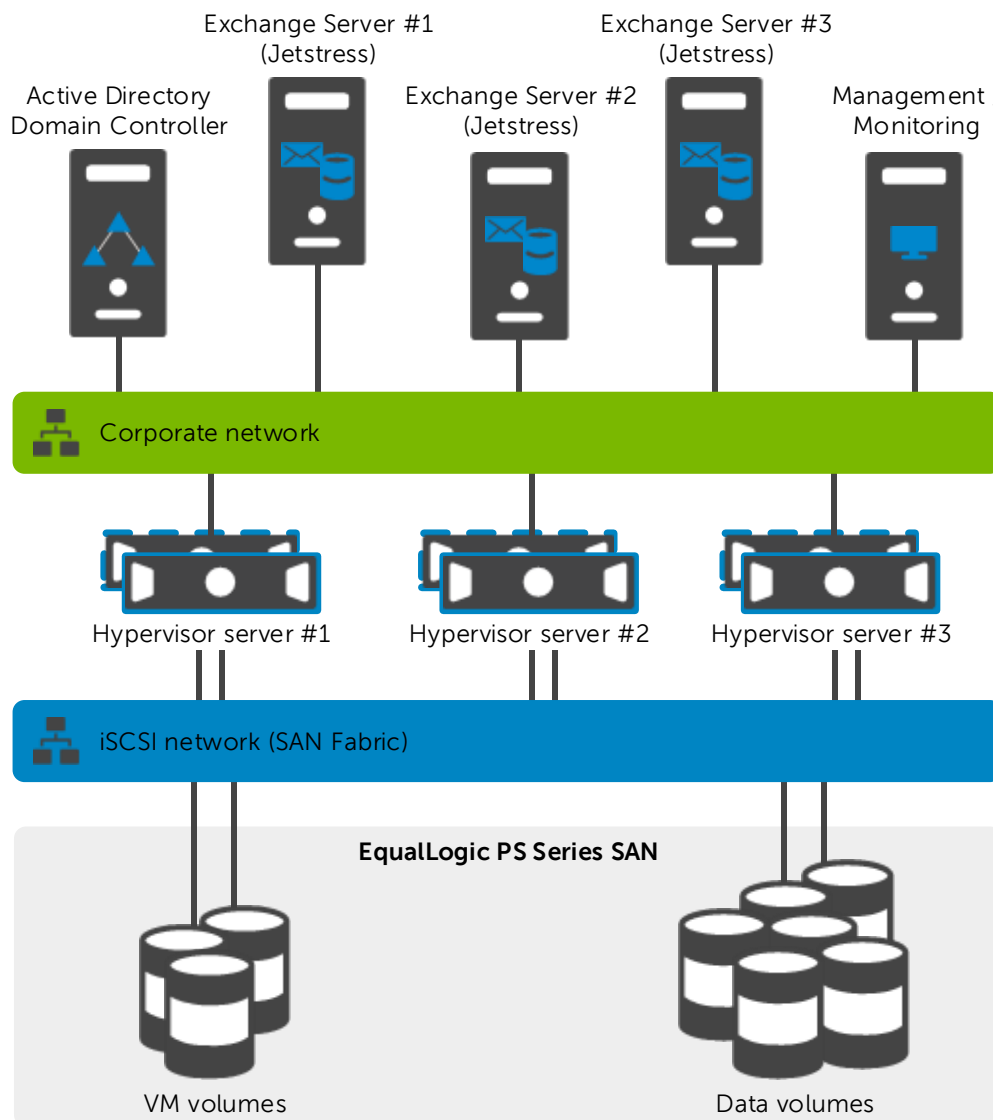
Figure 1   Conceptual system design for the solution components

The key elements of this design are:

- Single Active Directory forest, single domain, single site
- Centralized management and monitoring with dedicated resources (both physical and virtual)
- Separated network design to maintain traffic isolation between traditional LAN and iSCSI access
- Building block design approach for mailbox server with Jetstress

3.2   Physical system configuration

The physical components and the connections beneath the virtual infrastructure are shown in Figure 2.
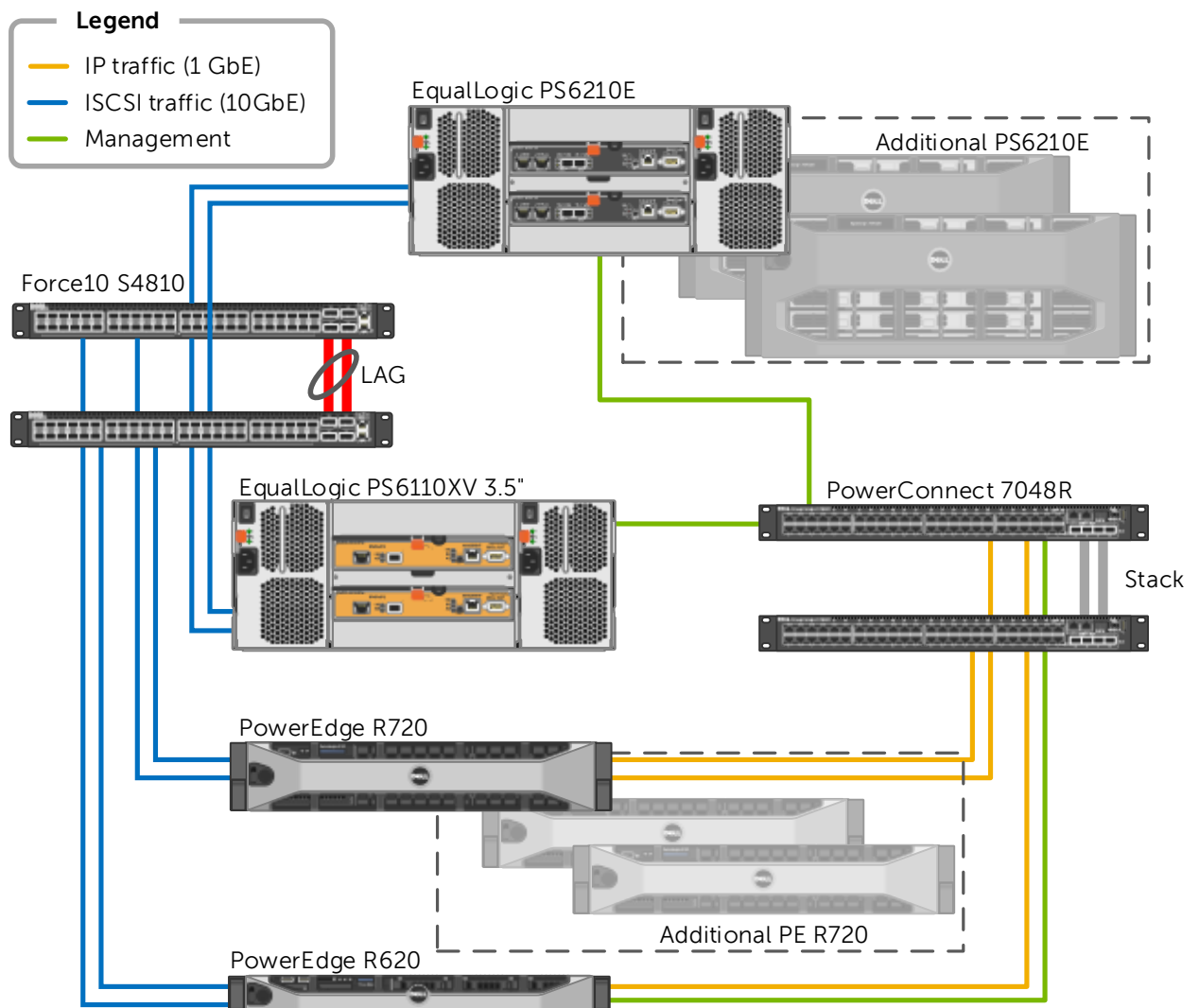
Figure 2   Physical system design for the components of the solution

The solution architecture was deployed on Dell rack servers with top of rack (ToR) Ethernet network switches dedicated to IP traffic (traditional client/server, management, and hypervisor communications) and to iSCSI storage access respectively. The hardware elements included in the architecture were:

- One to three PowerEdge R720 rack servers that powered the hypervisors beneath the simulated VMs
- One PowerEdge R620 rack server that powered the hypervisor beneath the centralized management and monitoring infrastructure
- EqualLogic iSCSI SAN provisioned with one PS6110XV 3.5" and up to three PS6210E arrays (10GbE)
- Two PowerConnect 7048R Ethernet switches to support LAN IP traffic configured in stack
- Two Force10 S4810 Ethernet switches to support the iSCSI traffic configured with Link Aggregation Group (LAG) consisting of two connections between the ToR S4810 switches

Note: Configuration details used in the solution infrastructure, including a hardware and software list, SAN array characteristics, hypervisor and VMs relationship, and physical and virtual network connections, are listed in Appendix A.

3.3 Storage and volume layout

The configuration of EqualLogic SAN arrays and the volumes underlying the Exchange databases included:

- One EqualLogic group, configured with one PS6110XV 3.5 inch array and a single pool
- One EqualLogic group, configured with up to three PS6210E arrays and with one to three pools associated with a one-to-one ratio with the array members
- RAID 6 policy applied as a reference configuration
- One volume allocated in the PS6110XV pool for each host to store the file images, configuration and temporary files for the corresponding hosted VMs
- A set of data volumes allocated within the PS6210E pools, dedicated to the Exchange databases and respectively assigned to the R720 hosts or to the Exchange VMs hosting the simulation
- Exchange mailbox databases and their private set of log files are deployed in the same volumes

Group #1 (PS6110XV)	Group #2 (PS6210E)
Volume #1 — VM	Volume #1 — Database and logs
Volume #2 — VM	Volume #2 — Database and logs
Volume #n — VM	Volume #n — Database and logs

Figure 3   Volumes and database/logs layout

4 Virtualization of Microsoft Exchange storage workload

Exchange Server is a product built on a wide set of components and services that cooperate to support the various requirements needed to design and deploy a messaging infrastructure with advanced capabilities in every organization. The latest Exchange Server version offers a simplified topology that reduces the number of server roles available while consolidating and optimizing the services in two layers. The front-end services deal with the clients (Client Access Service role), and the back-end services are in charge of data management and message transport (Mailbox role).

Because more capabilities can be now implemented in fewer layers, the first outcome is an increase in processing demand for the servers deployed to support Exchange Server 2013. This is a challenge that should be planned carefully from both capacity and performance standpoints and tailored to each organization in order to be successful with virtualized infrastructures.

The appropriate sizing of the storage subsystem is a key factor in the mailbox role performance and can contribute heavily to easing bottlenecks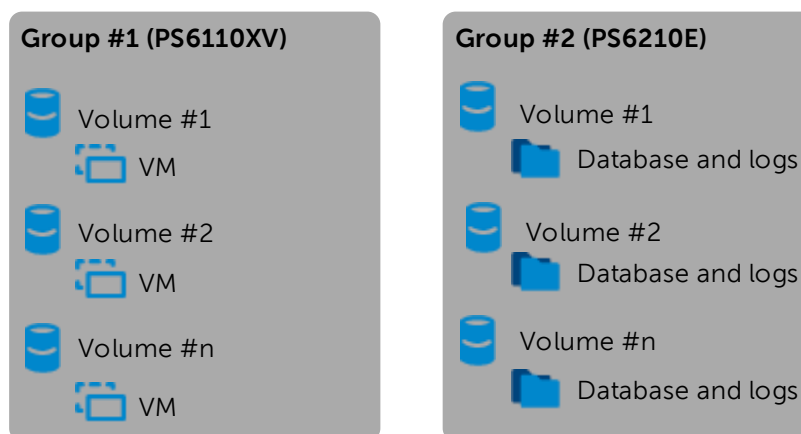 or administrative overhead for that layer. To better introduce the analysis available in the later sections of this paper, the components underneath the Managed Store of Microsoft Exchange Server 2013 are examined in section 4.1.

4.1 Exchange store elements

Accessing mailbox databases is the primary activity that generates I/O on an Exchange Server storage subsystem. A database is a logical representation of a collection of user or system mailboxes, and it is also an aggregation of files on the disk that are accessed and manipulated by a set of Exchange services following distinct rules (for example, the Information Store, Search feature or Replication Service).

Database file (*.edb): The container for user mailbox data. Its content, broken into database pages of 32 KB, is primarily read and written randomly as required by the Exchange services running on the mailbox server role. A database has a 1:1 ratio with its own *.edb database file. The maximum supported database size in Exchange Server 2013 is still 16 TB, where the Microsoft guidelines recommend a maximum 200 GB database file in a standalone configuration and 2 TB if the database participates in a replicated DAG environment.

Transaction logs (*.log): The container where all the transactions that occur on the database (such as create, modify and delete messages) are recorded. Each database owns a set of logs and keeps a one-to-many ratio with them. The logs are written to the disk sequentially, appending the content to the file. The logs are read only when they are in a replicated database configuration within a DAG or in the event of a recovery. The log truncation method is the process to remove old log files preventing them from accumulating. The truncation happens when a consistent backup is performed or when the log configuration is set to circular logging.

Checkpoint file (*.chk): A container for metadata tracking when the last flush of data from the memory cache to the database occurred. Its size is limited to 8 KB and, although repeatedly accessed, its overall amount of I/O is limited and can be ignored. The database keeps a 1:1 ratio with its own checkpoint file and positions it in the same folder location as the log files.

Search Catalog: A collection of flat files (content index files) built by the Microsoft Search Service, having several file extensions and residing in the same folder. The client applications connected to Exchange Server benefit from this catalog by performing faster searches based on indexes instead of full scans.

Exchange Server uses a proprietary format called Extensible Storage Engine (ESE) to access, manipulate, and save data to its own mailbox databases. This same format is employed on the Exchange HUB server role for the queue databases. ESE technology, previously known as Jet Database Engine, has evolved through several versions of Exchange Server and has been a part of several Microsoft products since its inception (for example, Microsoft Access, Active Directory, File Replication Service, WINS server, and Certificate Services).

The ESE is an Indexed Sequential Access Method (ISAM) technology that organizes database data in B-Tree structures. Ideally, these databases are populated by data kept together or adjacent. When this does not occur, external reorganization or defragmentation processes should be used to restore the optimal data contiguity in these structured databases.

To summarize, an Exchange mailbox database is subject to a subset of tasks performing storage access:

- The regular read and write access required to retrieve and store user mailbox data (according to the Exchange cache policy)
- The online defragmentation and compacting activities due to the B-Tree optimization
- The background database maintenance, which includes recoverable items cleanup, deleted mailboxes purge, and other activities addressing logical object support
- The checksum database scan to verify data block integrity (sequential read activity), which can be set to run constantly in the background or at a scheduled time

Furthermore, Exchange Server offers a specialized offline manual defragmentation task that runs while the database is dismounted by taking advantage of the ESEUTIL.EXE command line tool. The principal goal of this task is to reclaim the empty space left in a database by online defragmentation and to shrink the size of the *.edb file itself. This returns the free space to the operating system volume.

Note: It is not recommended to include offline defragmentation in a regular maintenance plan due to the disruption in the availability of the database, the rupture of the logs chain, and the need for database re-seeding in case of DAG configuration.

Exchange DAG is a pool of up to 16 networked servers that hosts multiple copies of the same Exchange database or databases where only one of the copies is active at a specific point-in-time within the group. The other copies are passive and contain data sourced from replicated and replayed transaction logs.

Log Checkpoint depth refers to the amount of logs written to the disk and that contain transactions not yet flushed to the database file. In Exchange Server 2013, during a DAG failover the database cache is no longer flushed, since it is treated as a persistent object. Therefore, the log checkpoint for the passive databases is increased to 100 to reduce the write I/O and to reduce the failover time since the passive database has to pre-read less data.

5 Validating the Exchange workload trends

Microsoft Exchange is a flexible deployment platform offering many different configurations and set of thresholds to who plans the solution and consequently to the end-user. The selection of some variables influences the performance and administration of the messaging infrastructure.

Since the focus of this paper is the behavior of scaling the messaging workload and on the technologies supporting it, the following variables were kept constant at the most common averages during the tests to increase the consistency of the results. This included user profile, mailbox size, number of mailbox ratio per database, and database size.

Note: the results presented in this white paper are gathered from simulations executed in a lab built to the listed specifications. Other generally available tools or sizing calculators might have results slightly different based on the amount of assumptions made from each particular tool.

The reference Exchange server deployment used is detailed in Table 1. For each scenario evaluated, an explicit description of the relevant differences is reported in a corresponding section of this paper.

Table 1   Reference configuration for Microsoft Jetstress 2013 tests

Reference configuration: factors under study	
Number of simulated users / mailboxes	5,000 concurrent users
Number of databases	10 databases (active)
iSCSI initiator	software initiator (Windows Server 2012 R2)
Array model, SAN configuration	one PS6210E (24x 4TB NL-SAS 7.2K disks), one pool
Reference configuration: consistent factors across each scenario	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.134 IOPS (with DAG)
Mailbox size	2 GB each
Mailbox allocation per database	500 mailboxes per each mailbox database
Database size	1 TB each
Number of database replica copies	2 (two node DAG)
Background database maintenance	Enabled
Windows Disk/Partition File System	Basic disk, GPT partition, default alignment NTFS, 64KB allocation unit size
RAID policy	RAID 6
Test duration	2 hours + time required to complete DBs checksum

Note: For a comprehensive list of storage configuration options and supported scenarios for Exchange Server, refer to the article "Exchange 2013 Storage Configuration Options" available at: http://technet.microsoft.com/en-us/library/ee832792%28v=exchg.150%29.aspx

The following list provides the metrics and pass/fail criteria recorded while completing the tests. Most of this information is outlined by the Jetstress tool report or is verified through the recording of Windows Performance Monitor and Dell EqualLogic SAN Headquarters counters. Microsoft thresholds for Exchange Server storage validation are reported as well.

Database Reads Average Latency (ms) is the average length of time to wait for a database read operation (random reads). It should be less than 20 ms according to Microsoft threshold criteria.

Database Writes Average Latency (ms) is the average length of time to wait for a database write operation (random writes). It should be less than 20 ms according to Microsoft threshold criteria.

Logs Writes Average Latency (ms) is the average length in time to wait for a log file write operation (sequential writes). It should be less than 10 ms according to Microsoft threshold criteria.

Planned Transactional IOPS are the target amount of IOPS for the test (calculated by multiplying the number of users by the IOPS per mailbox).

Achieved Transactional IOPS are the amount of IOPS actually performed by the storage subsystem to address the transactional requests. The result should not diverge more than 5% from the planned IOPS to be considered a successful test iteration according to Microsoft Jetstress.

LOGs IOPS are the IOPS performed against the log files during the transactional test. They are not directly taken into account as part of the transactional IOPS, but are tracked separately instead.

Additional IOPS are the IOPS generated for the database (DB) maintenance, log files replication and all the remaining activities on the storage subsystem, calculated as the difference between the IOPS provisioned by the EqualLogic SAN and the previously reported transactional and logs IOPS.

Total IOPS of the SAN is the sum of achieved transactional IOPS, Logs IOPS and additional IOPS. It represents the entire IOPS footprint performed against the back-end SAN during a test. It is recorded at the SAN level and verified with the Exchange host.

Note: For details about the simulation tool, Microsoft Jetstress 2013, refer to Appendix B.

5.1 Characterize the impact of the iSCSI initiator collocation

In networked storage systems, iSCSI communications happen through the wire between two layers or end nodes. The iSCSI initiator functions as a client, which accesses the storage resources located on the target. The target acts as a server hosting the data. Messages and commands, similar to SCSI, are sent over the IP network between the two end nodes.

A software initiator is an implementation of these functionalities by code. It usually runs as a device driver and reuses the network cards available in the operating system to emulate SCSI devices. While implementing a virtual infrastructure, a choice for the connectivity with the EqualLogic SAN through a software based iSCSI initiator would be the collocation of the initiator itself.

- Guest initiator: Software located and running on the guest VMs, which allows direct connections to the volumes residing on the SAN through the virtual network adapters of the guest. The settings of the VMs include additional virtual network adapters dedicated to SAN traffic. The host hypervisor is not aware of the type of traffic traversing the virtual network adapters.
- Host initiator: Software located and running on the host hypervisor, which allows connection to the volumes residing on the EqualLogic SAN from the Windows Hyper-V host through physical network adapters dedicated to the SAN traffic. VHDX files are created on the SAN volumes attached to the host and then added as SCSI disks to the settings of the VMs. The VMs are not aware if disks reside on either local storage or the SAN.

The goal of this analysis is to establish the Exchange KPI trends and IOPS ratios when comparing a deployment implemented through host iSCSI initiator against one implemented with a guest initiator in the reference environment defined previously. Table 2 shows a summary of the configurations used for this scenario (reference and variations).

Table 2   Test parameters: iSCSI software initiator collocation

Reference configuration: factors under study	
iSCSI initiator software collocation	Host SW initiator (Microsoft Windows Server 2012 R2 with Hyper-V)
	Guest initiator (Microsoft Windows Server 2012 R2)
Reference configuration: consistent factors within this scenario	
Messages per day per mailbox / IOPS per mailbox	200 messages/0.134 IOPS (with DAG)
Number of simulated users / mailboxes	5,000 concurrent users
Mailbox size	2 GB each
Number of databases	10 databases (active)
Mailbox allocation per database	500 mailboxes per each mailbox database
Database size	1 TB each
Number of database replica copies	2 (two node DAG)
Array model, SAN configuration	one PS6210E (24x 4TB NL-SAS 7.2K disks), one pool
RAID policy	RAID 6

The outcome of this comparison shows that both deployments performed closely to each other for both the amount of IOPS delivered and the average database read latencies recorded. The only noticeable difference is in the write latencies where the database latency is lower in a host initiator environment while the log files latency is lower in the guest initiator deployment. The VHDX format performance was comparable to direct connected volumes and demonstrates its best with random access patters. The direct connections through the guest VMs, while still under the warning thresholds, performs better with sequential access streams of data.

Figure 4 shows the results collected from the two instances of Exchange Jetstress simulations.
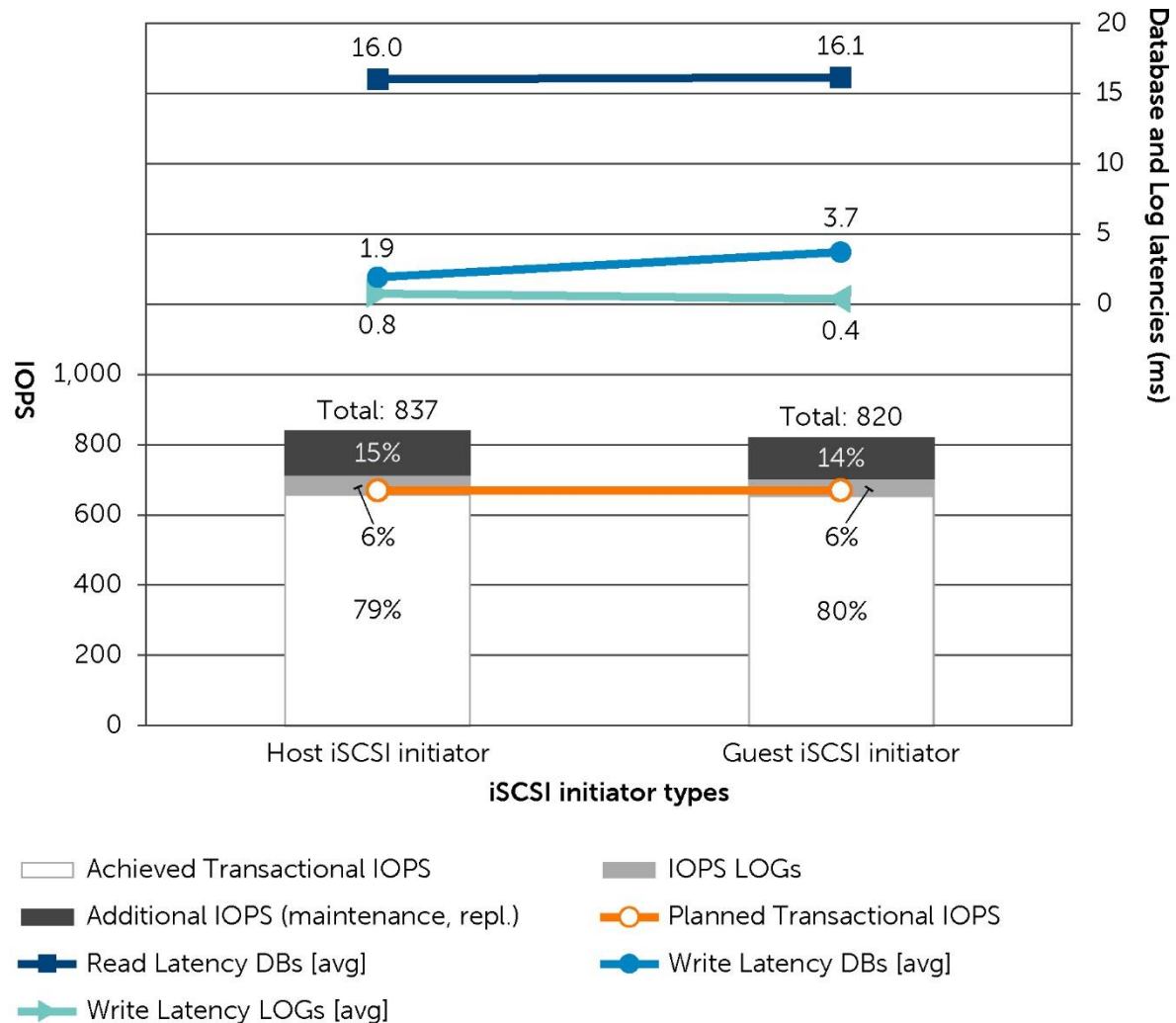

Figure 4   Storage trend with different iSCSI initiator collocation with 5,000 users evenly distributed across ten replicated databases

Note: The numerical results presented in this paper were rounded to one decimal place for latencies, and whole numbers for percentages and IOPS. The calculation to obtain percentages was executed with the raw values gathered at three decimal places.

5.2 Scale up the user count per storage unit

The mailbox count represents the total number of mailbox users hosted by the Exchange server and stored on the corresponding SAN volumes. As the user count increases, the capacity allocated and usually the workload generated raise, with the assumption of concurrent users. Meanwhile, the hosting of a large number of mailboxes by a single server increases the potential risk of widely distributed loss of service in case of server failure, and demands a highly available messaging solution.

The following analysis establishes the Exchange KPI trends and IOPS ratios when increasing the number of concurrent users and the global workload on the SAN in the reference environment defined previously. Table 3 shows a summary of the configurations used for this scenario (reference and variations).

Table 3   Test parameters: variable number of concurrent mailbox users

Reference configuration: factors under study (Series #1)	
Number of simulated users / mailboxes	4,000 / 5,000 / 6,000 concurrent users
Number of databases	8 / 10 / 12 databases (active)
iSCSI initiator software collocation	Host SW initiator (Microsoft Windows Server 2012 R2 with Hyper-V)
Reference configuration: factors under study (Series #2)	
Number of simulated users / mailboxes	4,000 / 4,500 / 5,000 / 5,500 concurrent users
Number of databases	8 / 9 / 10 / 11 databases (active)
iSCSI initiator software collocation	Guest initiator (Microsoft Windows Server 2012 R2)
Reference configuration: consistent factors within this scenario	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.134 IOPS (with DAG)
Mailbox size	2 GB each
Mailbox allocation per database	500 mailboxes per each mailbox database
Database size	1 TB each
Number of database replica copies	2 (two node DAG)
Array model, SAN configuration	one PS6210E (24x 4TB NL-SAS 7.2K disks), one pool
RAID policy	RAID 6

The provision of mailbox users by a gradual and predefined number represents a valuable simplification for the administration of the underling infrastructure, which can be then organized to conveniently spin resources by predefined increments, both in average workload and capacity.

The two series of tests used the same building block to grow the user count: databases of 1 TB hosting 500 users each and deployed on their own SAN volumes. First, the deployment with host initiator was assessed against three progressively increasing workloads scaling from 4,000 to 6,000 concurrent mailbox users with an intermediate step at 5,000 users. Then, the deployment with guest initiator was confronted with four workloads of 4,000, 4,500, 5,000 and 5,500 concurrent mailbox users.

Figure 5 shows the results collected from the three instances of Exchange Jetstress simulations of Series 1.
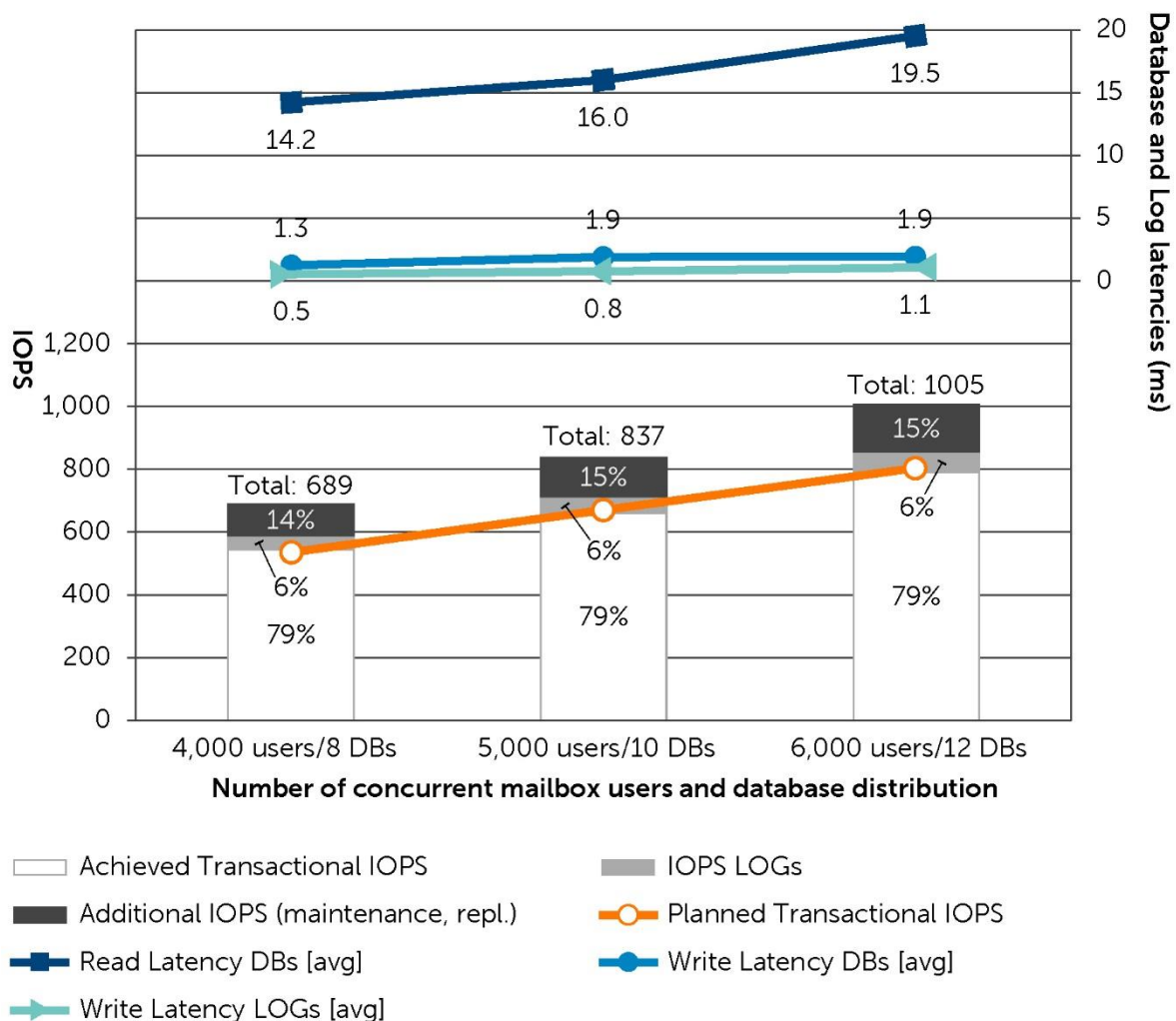
Figure 5   Storage trend while scaling up the number of concurrent mailbox users increasing the number of DBs (configured with host initiator)

The outcome of the first series reveals a near linear scalability on the entire set of validation, and leaves all the results under the predefined thresholds set on databases and log file latencies. From the chart in Figure 5, the database read latency deteriorated faster with the increase of workload, but the real trend is revealed by Table 4 where the workload increased by 25% and then 50% were never reached by the DBs

read latency rise. The write latencies rose faster than the workload increase, but, since the absolute value was extremely low. This rise is worth noticing, but did not have an impact on the overall result.

Table 4   Percentage comparison of the increase in workload, IOPS and latencies for Series #1

Number of concurrent users	Total IOPS	R latency DBs	W latency DBs	W latency Logs
4,000 or base workload of 100%	689 or 100%	14.2ms or 100%	1.3ms or 100%	0.5ms or 100%
5,000 or workload of 125%	837 or 121%	16.0ms or 113%	1.9ms or 154%	0.8ms or 143%
6,000 or workload of 150%	1005 or 146%	19.5ms or 137%	1.9ms or 154%	1.1ms or 201%

The outcome of the second series shows an extremely close trend to the one provided by the first series. This trend is apparent in the values that fell under the Exchange latency thresholds at every point. Once again, the chart lines suggest a steeper increase of the database read latency at the heaviest workload, but the numerical percentage reported in Table 5 proves the read latency growth was linear with the workload increase, as well as the log files write latency. Instead, the database write latency rose faster than the workload.

Table 5   Percentage comparison of the increase in workload, IOPS and latencies for Series #2

Number of concurrent users	Total IOPS	R latency DBs	W latency DBs	W latency Logs
4,000 or base workload of 100%	680 or 100%	14.4ms or 100%	2.3ms or 100%	0.3ms or 100%
4,500 or workload of 112.5%	760 or 112%	15.4ms or 107%	3.0ms or 128%	0.4ms or 113%
5,000 or workload of 125%	820 or 121%	16.1ms or 112%	3.7ms or 161%	0.4ms or 114%
5,500 or workload of 137.5%	894 or 131%	18.7ms or 130%	4.4ms or 190%	0.4ms or 126%

Figure 6 shows the results collected from the four instances of Exchange Jetstress simulations of Series 2.
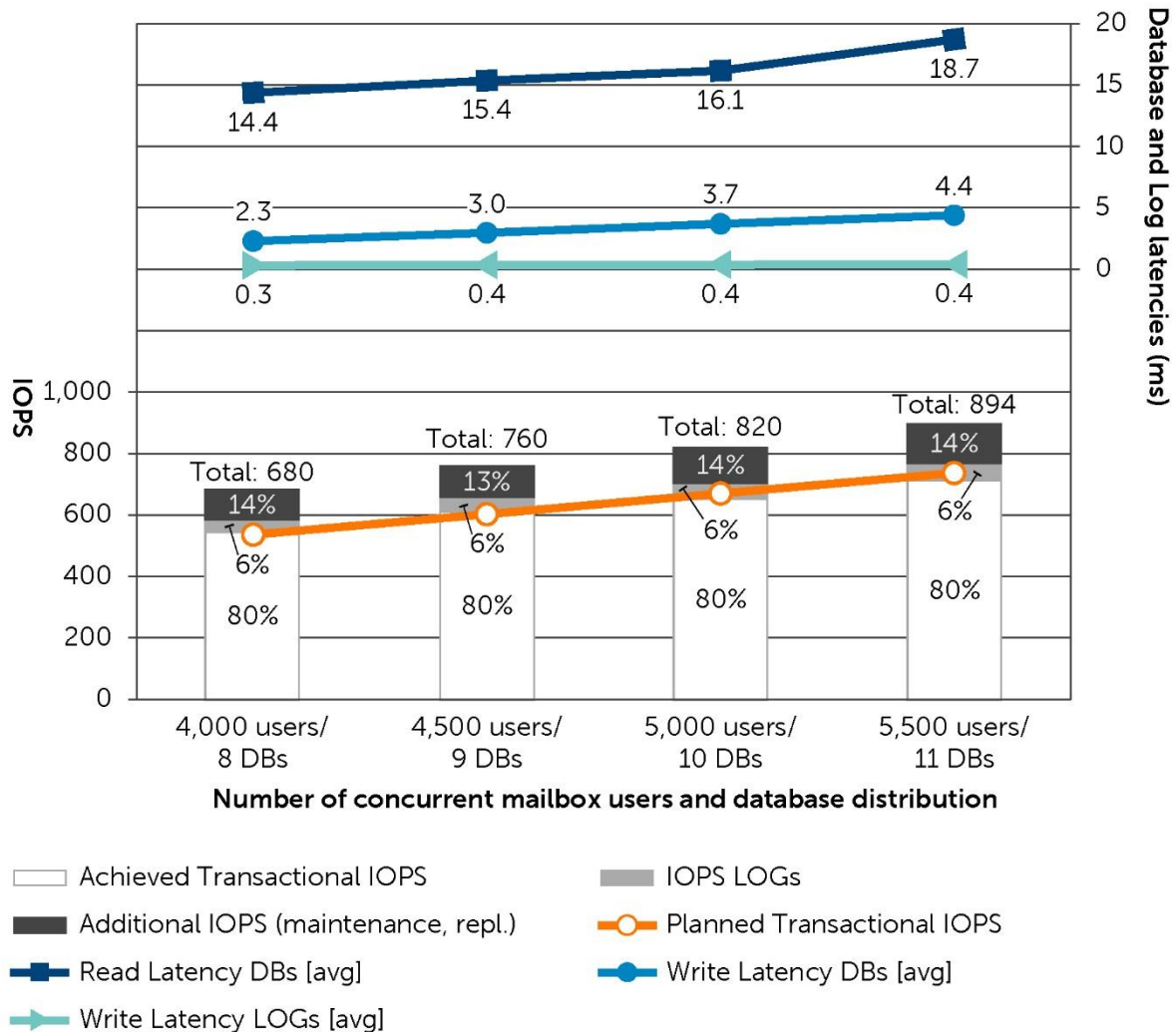
Figure 6   Storage trend while scaling up the number of concurrent mailbox users increasing the number of DBs (configured with guest initiator)

The summary of the two series underlines both deployment methodologies (host or guest initiator) are suitable for this workload and can scale linearly. When compared to each other, they show distinct characteristics (see Section 5.1). From a pure performance scalability point of view, the host initiator configuration validated provides a higher resiliency to database read latency. This marks the first threshold for a comparable number of concurrent mailbox users. The guest initiator, while having the same results for average workloads, rose quicker that the host initiator when driven close to the boundaries.

Selecting the most appropriate iSCSI initiator in a deployment is driven not only by performance factors, but also by considerations on data protection, management and administration, and the standardization of virtualized infrastructures. Each of these subjects should carefully be weighed and measured against the business objectives of the organization where the messaging infrastructure is being planned in order to take the most appropriate decision between these two options.

Note: For additional information about Exchange data protection options with Dell EqualLogic SAN, refer to *Best Practices for Enhancing Microsoft Exchange Server 2010 Data Protection and Availability using Dell EqualLogic Snapshots*, available at: http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19961113/download.aspx.

5.3   Scale out the SAN

The storage environment created by EqualLogic allows a SAN to be easily scaled out on an as-needed basis by adding additional array members to the group of the arrays already deployed. It is a non-disruptive task for the preexisting data, able to deliver additional capacity and performance when required.

The following analysis established the Exchange KPI trends and IOPS ratios when scaling out the SAN by incrementing the number of arrays in order to support an increased number of concurrent users. Table 6 shows a summary of the configurations used for this scenario (reference and variations).

Table 6   Test parameters: Scale out the SAN while scaling up the number of concurrent users

Reference configuration: factors under study (Series #1)	
Number of simulated users / mailboxes	4,000 / 8,000 / 12,000 concurrent users
Number of databases	8 / 16 / 24 databases (active)
Array model, SAN configuration	1x / 2x / 3x PS6210E (24x 4TB NL-SAS 7.2K disks) 1x / 2x / 3x pools (one array per each pool)
Reference configuration: factors under study (Series #2)	
Number of simulated users / mailboxes	5,000 / 10,000 / 15,000 concurrent users
Number of databases	10 / 20 / 30 databases (active)
Array model, SAN configuration	1x / 2x / 3x PS6210E (24x 4TB NL-SAS 7.2K disks) 1x / 2x / 3x pools (one array per each pool)
Reference configuration: consistent factors within this scenario	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.134 IOPS (with DAG)
Mailbox size	2 GB each
Mailbox allocation per database	500 mailboxes per each mailbox database
Database size	1 TB each
Number of database replica copies	2 (two node DAG)
iSCSI initiator software collocation	Host SW initiator (Microsoft Windows Server 2012 R2 with Hyper-V)
RAID policy	RAID 6

The building block used to scale out the SAN for larger environments was framed around the same user per database ratio of 500 to 1 defined previously. The workload was then doubled and tripled by adding infrastructure blocks. The components involved in the infrastructure building block were one EqualLogic PS6210E array (each configured in an independent pool), one Windows Hyper-V host and one VM running Exchange mailbox server (simulated by Jetstress).

The two series of tests referenced previous workloads as a starting point. The first progression began at 4,000 concurrent mailbox users, and then scaled to 8,000 and 12,000 users. The second, more intensive, progression moved from 5,000 to 10,000 and then 15,000 concurrent mailbox users. Figure 7 shows the results collected from the six instances of Exchange Jetstress simulations of Series 1 and Series 2 placed side by side.

Overall, the outcome of this study acknowledges the ability of the EqualLogic SAN to deliver linear performances from one to three arrays in the same group while keeping all Exchange KPI within the recommended thresholds.

The first series based on the 4,000 user building block demonstrated a remarkable yet completely predictable behavior. It provided a solution with room for future growth on both capacity and performance on the single array, as well as a building block that can be adopted with confidence for horizontal scalability deployment.

The second series, based on a more demanding workload of 5,000 users per single array, still performed with good results. There was only a minor degradation in the database read latency during the horizontal scaling exercise of 10% for two arrays, and then 15% for the three arrays.
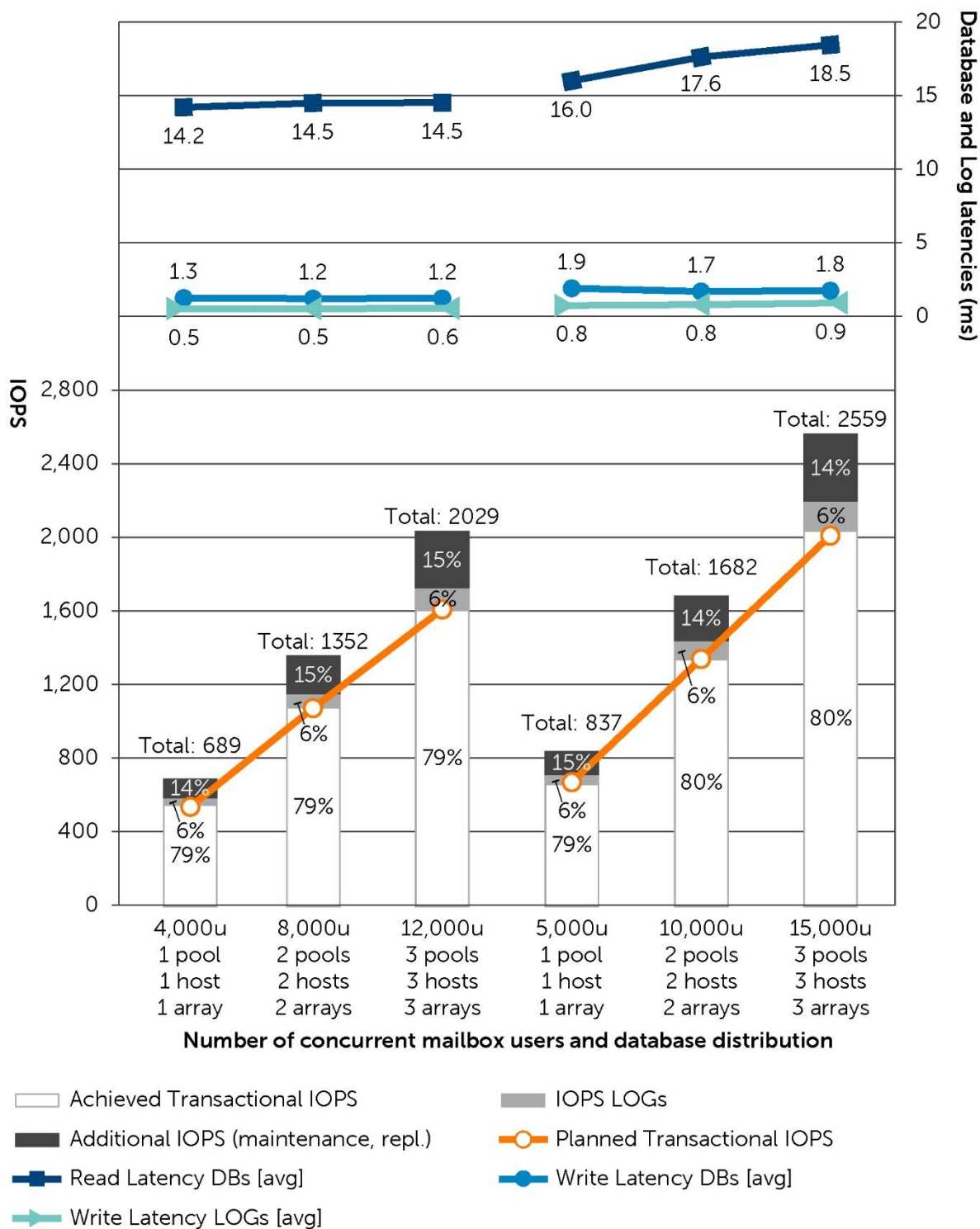
Figure 7   Storage trend scaling out the SAN while scaling up the number of users

6   Best practice recommendations

Refer to these best practices to plan and configure EqualLogic arrays, Exchange Server 2013 and VMs.

Storage best practices

- Distribute network port connections on the controllers according to the port failover guidelines and the redundancy implemented on the network switches.
- Maintain as close as possible a 1:1 ratio between the number of network ports on the active controller of the array and the number of host network adapters to maximize the utilization of the available bandwidth and to minimize oversubscription.
- Use a Multipath I/O (MPIO) extension provided by EqualLogic to improve iSCSI performance and reliability. This is a Device Specific Module (DSM) in Windows provided by the EqualLogic HIT kit.
- Choose the most appropriate RAID policy when designing the environment according to the performance, capacity, and tolerance to failure requirements of the environment.
- For environments with a very large number of volumes attached to the servers, monitor the amount of iSCSI connections per host generated.

Network best practices

- Design separated network infrastructures to isolate the LAN traffic from the SAN traffic (iSCSI).
- Secure the networks, including the SAN, to protect resources from external threats such unwanted packet sniffing or connection hijacking. Security mechanisms should include physical security, authentication and application level encryption.
- Implement redundant components (switches, ISLs, and network adapters) to provision a resilient infrastructure between the endpoints (stack, LAG, load balancing, or network card teaming).
- Disable spanning tree for the switch ports hosting the PS Series array controller connections and enable Rapid Spanning Tree Protocol instead.
- Enable flow control for the switch ports hosting the PS Series array controller connections.
- Enable flow control on the host network adapters and on the VM virtual network adapters connected to SAN traffic (iSCSI).
- Enable jumbo frames (large MTU) for the switch ports hosting the PS Series array controller connections, on the host network adapters and on the VM virtual network adapters connected to SAN traffic (iSCSI).
- Evaluate jumbo frames (large MTU) for the LAN network when appropriate (limited by the type of devices the traffic traverses).
- Strongly consider SR-IOV and VMQ technologies if supported by the host network adapters connected to the SAN traffic (iSCSI). Evaluate them for the LAN network too when appropriate. Enable these options through the virtual stack on the virtual switches and virtual network adapters.
- Enable Large Send Offload, IP, TCP, and UDP Checksum Offload for both RX and TX on the host network adapters and on the VM virtual network adapters connected to the SAN traffic (iSCSI).
- Maximize the size and number of buffers for both transmit and receive on the host network adapters and on the VM virtual network adapters connected to the SAN traffic (iSCSI).

Hypervisor and VMs best practices

- Install a Windows Server Core version in the root partition of the Hyper-V role server to reduce the software maintenance, the attack surface, the memory, and disk space footprint. Otherwise, when installing a traditional Windows Server with Hyper-V technology with the GUI, minimize the use of additional software, components and/or roles in the root partition.
- The use of NUMA is advised to address the management of VMs with large and very large memory settings. Verify the number of NUMA nodes available in the system, based on the number of processors, and then design and size the VMs to have their memory resources entirely contained in a single NUMA node when possible. Spanning a VM memory across multiple NUMA nodes can result in less efficient usage of the memory and can decrease performance.
- Exchange Server 2013 is a memory intensive workload. Configure static memory in the settings of each VM to avoid possible resources contention created by the dynamic memory management and to comply with the current support directives on this matter.
- Investigate the opportunity to reserve the resources assigned to the Exchange VMs by using the resource control settings.
- Plan the reserve size for the volumes hosting the hard disk files of the VMs (*.vhdx, *.vhd), and include the space required for memory image files (*.bin), saved states (*.vsv), or snapshot files (*.avhdx, *.avhd).
- Remember the supportability constraints of Exchange Server 2013 when deployed in a virtualized environment around suspended state, VM snapshots and replicas, and differencing disk image files.
- While performance of dynamically expanding disks has improved, fixed disks are preferred to deploy production environment VMs due to the risk of elevated fragmentation or high latency while disk expansion occurs.
- Isolate the host management traffic from the VM traffic by using virtual switches not enabled for management.
- Avoid mixing LAN and iSCSI traffic on the same virtual network adapters by designing isolated LAN and iSCSI networks.
- Configure a dedicated virtual switch for each network adapter connected to the SAN traffic (iSCSI).

Exchange Server installation best practices

- Use Basic disk type for all EqualLogic volumes.
- Use GUID partition table (GPT) for Exchange volumes.
- Use default disk alignment provided by Windows 2008 or greater.
- Use NTFS file system with 64 KB allocation unit for Exchange database and log partitions.
- Evaluate the use of mount points for all the SAN volumes or the attached virtual disks to increase management flexibility and database portability. Mount points are required when the number of volumes is greater than the number of available drive letters in the servers.
- Prevent Windows Server from assigning drive letters to volumes by disabling the auto-mount option to minimize unwanted volume letter assignment in a mount point managed environment.
- Deploy Windows operating system and Exchange data in physically separated disk drives, or in separated pools for an EqualLogic storage environment.
- Database and log file isolation is not required when deployed in a DAG environment.

- Leave background database maintenance (BDM) enabled (24x7) and account for the additional load. The BDM is activated by default on every replica copy of a DAG configuration.
- Do not share the disk drives for active and replicated copies of an Exchange mailbox database in a DAG environment. If there is a failure of a set of drives with multiple copies of the same data, the resilience or the perceived availability of the applications would be affected. Dedicate separate EqualLogic pools for each replicated instance of the data instead.

Know the workload

- Do not begin a deployment without having a solid understanding of the current messaging workload. In the case of a new deployment, collect estimates based on business cases matching your organization size and drive conservative figures for the average user profiles. The use of Microsoft Exchange 2013 Server role calculator is advised to learn basic estimates for new or changing messaging environments.
- Forecast the workload gap between the current version of Exchange Server and Exchange Server 2013 when planning to design a storage solution jointly with a migration.

Distribution of databases and user mailbox count per database

- Carefully plan the number of databases and the mailbox count per database in the entire organization. The right balance of the number of mailbox databases to support a defined number of users is mostly based on administrative policies.
- Larger databases fitted with a high number of users have a smaller impact on the storage subsystem, but provide less flexibility because of large database files maintenance and replication.
- Reduced mailbox count per database provides a more agile environment when using a traditional backup application, or when administrative tasks require temporarily dismounting a database causing downtime. Deploying many users in a single database is perceived as a single point of failure or as a constraint.
- Microsoft Windows operating system supports a precise maximum amount of iSCSI targets and connections. When a guest initiator is in use together with a great number of volumes, monitor the number of connections and do not pass the limits to avoid an unsupported deployment scenario.

Growth and overhead

- Understand the wider business requirements and company strategy to predict the organization growth instead of following it.
- Design an environment based on building blocks that are easily replicable to address sudden changes in business requirements.
- Always account for at least 20% overhead of the maximum sizes calculated to address unforeseen growth.

A Configuration details

A.1 Hardware components

Table 7 lists the details of the hardware components used for the configuration.

Table 7   Hardware components

Component	Description
Servers	Dell PowerEdge R620 server, Firmware 2.2.2 • 2x Eight Core Intel Xeon E5-2665 Processors, 2.4 Ghz, 20M Cache • RAM 32 GB (4x 8GB) • iDRAC7 Enterprise, Firmware 1.56.55 • PERC H710 Mini RAID controller, Firmware 21.2.0-0007 • 4x 146 GB 15K SAS (2x RAID-1, stripe 1MB) • 4x Broadcom NetXtreme 5720 Quad Port 1GbE Base-T onboard, Firmware 7.8.53 • 2x Broadcom NetXtreme II 57810 Dual Port 10GbE Base-T, Firmware 7.8.53 Dell PowerEdge R720 servers, Firmware 2.0.19 • 2x Eight Core Intel Xeon E5-2665 Processors, 2.4 Ghz, 20M Cache • RAM 384 GB (24x 16GB) • iDRAC-7 Enterprise, Firmware 1. 56.55 • PERC H710 Mini RAID controller, Firmware 21.2.0-0007 • 4x 146 GB 15K SAS (2x RAID-1, stripe 1MB) • 4x Broadcom NetXtreme 5720 Quad Port 1GbE Base-T onboard, Firmware 7.8.53 • 2x Broadcom NetXtreme II 57810 Dual Port 10GbE Base-T, Firmware 7.8.53
Network	2x Dell Force10 S4810 Ethernet switches, Firmware 9.2.0.0 • 48x 10GbE interfaces • 4x 40GbE interfaces • Installed top of the rack • Connected by 2x 40GbE redundant uplinks (LAG) 2x Dell PowerConnect 7048R Ethernet switch, Firmware 5.1.1.7 • 48x 1GbE interfaces • 4x 10GbE interfaces • Installed top of the rack • Connected by 2x16GbE redundant uplinks (STACK)
Storage	1 to 3x Dell EqualLogic PS6210E • Storage Array Firmware 7.0.3 • Dual 1 port 10GbE controllers • Dual 1 port 1GbE management interface • 24x 4TB 7.2K 3.5" NL-SAS disk drives, raw capacity 96 TB 1x Dell EqualLogic PS6110XV 3.5" • Storage Array Firmware 7.0.3 • Dual 1 port 10GbE controllers • Dual 1 port 1GbE management interface • 24x 600GB 15K 3.5" SAS disk drives, raw capacity 14.4 TB

A.2 Software components

The environment required to perform the simulations described in this paper included the following software components:

- Hypervisor: Windows Server 2012 R2 with Hyper-V on every physical host
- Dell OpenManage Server Administrator on every physical host
- Broadcom Advanced Control Suite on every physical host
- Operating System: Windows Server 2012 or 2008 R2 on every VM
- Operating System: Windows Server 2012 R2 on workload simulation VM and on management VM
- Dell EqualLogic Multi-pathing Extension Module to provide Dell MPIO access to the back-end SAN on the hypervisor directly accessing the SAN (host initiator scenarios)
- Dell EqualLogic Host integration Toolkit to provide Dell MPIO access to the back-end SAN on the VM directly accessing the SAN (guest initiator scenario)
- Dell EqualLogic SAN Headquarters to monitor the health and performance of the SAN
- Microsoft Exchange Jetstress to simulate the access to the storage subsystem from the mailboxes store simulated VM

The following software components are installed and configured to simplify the management of the environment and to support the failover cluster configuration, while they were not strictly required to accomplish the tests:

- Active Directory Domain Services and DNS Server roles for the domain controller VM
- Microsoft SCVMM for the management VM (not strictly required to accomplish the tests)

Table 8 lists the details of the software components used for the configuration.

Table 8   Software components

Component	Description
Operating Systems	Host servers: • Microsoft Windows Server 2012 R2 Datacenter Edition (build 9600) with Hyper-V • Dell OpenManage Server Administrator 7.4.0 • Broadcom Advanced control Suite 4 (version 16.3.6.0) • Dell EqualLogic Host Integration Toolkit 4.7.0 (host initiator connectivity only) • MPIO enabled using EqualLogic DSM for Windows Guest VMs: • Microsoft Windows Server 2012 R2 Datacenter Edition (build 9600) • Dell EqualLogic Host Integration Toolkit 4.7.0 (only for guest initiator connectivity) • MPIO enabled using EqualLogic DSM for Windows
Applications	Microsoft System Center 2012 Virtual Machine Manager Service Pack 1 (version 3.1.6011.0)

Component	Description
Monitoring tools	Dell EqualLogic SAN Headquarters 3.0 (build 3.0.0.7189) Microsoft Performance Monitor from the Windows Operating System
Simulation tools	Microsoft Exchange Jetstress 2013 (build 15.00.0775.000) • Exchange 2013 Server Database Storage Engine and Library SP1 (build 15.00.0847.030)

A.3   Network configuration

Two physical networks provide full isolation between regular IP traffic and iSCSI data storage traffic. Also, each regular IP network is segregated from the others by VLANs with tagged traffic. In order to achieve network resiliency to hardware faults, at least two physical switches are linked for the iSCSI data storage network using redundant uplinks (LAG) between the switches. Some relevant configuration aspects are:

- Flow control enabled for every port on S4810 switches
- Rapid Spanning Tree Protocol enabled for every edge port on S4810 switches
- Jumbo frames enabled for every port on S4810 and 7048R switches

Table 9 and Table 10 summarize the aspects of the physical and logical networks implemented in the reference architecture, and their purpose.

Table 9   Network configuration: network switch, purpose, and networks

Network Switch	Placement	Purpose	VLAN ID
PowerConnect 7048R	Top of rack	IP Traffic - LAN	100
		IP Traffic - Management	200
Force10 S4810 #1	Top of rack	iSCSI data storage traffic	1 (default)
Force10 S4810 #2	Top of rack	iSCSI data storage traffic	1 (default)

Table 10   Network configuration: host to switch connections

Server	Interface	NIC port	Purpose	
PowerEdge R620	BCM5720 #1	1x 1GbE	LAN and Management	
	BCM5720 #2	1x 1GbE	LAN and Management	IP traffic
	BCM5720 #3	1x 1GbE	VMs traffic	
	BCM5720 #4	1x 1GbE	VMs traffic	
	BCM57810 #1	1x 10GbE		

Server	Interface	NIC port	Purpose	
	BCM57810 #2	1x 10GbE	SAN management and monitoring (guest VM)	iSCSI data storage traffic
	Total ports = 6 (4 onboard 1GbE, 2 slot 10GbE)			
PowerEdge R720	BCM5720 #1	1x 1GbE	LAN and Management	IP traffic
	BCM5720 #2	1x 1GbE	LAN and Management	
	BCM5720 #3	1x 1GbE	VMs traffic	
	BCM5720 #4	1x 1GbE	VMs traffic	
	BCM57810 #1	1x 10GbE	SAN access (host initiator)	iSCSI data storage traffic
	BCM57810 #2	1x 10GbE		
	BCM57810 #3	1x 10GbE	SAN access (guest initiator)	
	BCM57810 #4	1x 10GbE		
	Total ports = 8 (4 onboard 1GbE, 4 slot 10GbE)			

A.4 Host hypervisor and VMs configuration

A virtual infrastructure built on Windows Server with Hyper-V hosted all the components of the test infrastructure. The primary elements of the virtual infrastructure configuration are:

- Windows Server 2012 R2 with Hyper-V deployed on all hosts, managed by the Hyper-V Role Administration tools or centrally by the SCVMM server
- Up to three identical hypervisor hosts (R720s) configured as member server of the domain
- One hypervisor host (R620) configured as member server of the domain
- All guests deployed from one image template of Windows Server 2012 R2 operating systems
- Host iSCSI initiator to access the system disks (VHDX) of the VMs, which reside on the EqualLogic PS6110XV array
- Host or guest iSCSI initiator, depending on the appropriate scenario, used to access the volumes containing data and hosted on the EqualLogic PS6210E arrays

Table 11 lists the relation between each hypervisor host and its respective set of VMs, with a brief summary of the virtual resources allocated for each VM.

Table 11   Configuration: guest to host placement

Host	VM	Purpose	vCPU	Memory	Storage	Network adapters
R620	DC01	Active Directory Domain Controller	2	4GB	250GB VHDX	2x VMBus Net Adapters (NIC teaming with LBFO)
	SCVMM01	System Center Virtual Machine Manager	4	8GB	250GB VHDX	2x VMBus Net Adapters (NIC teaming with LBFO)
R720s	JET#n	Exchange Server mailbox server simulation (Jetstress)	4	64GB	250GB VHDX + Exchange data volumes	2x VMBus Net Adapters (NIC teaming with LBFO)
						2x VMBus Net Adapters (iSCSI traffic)

Guest VMs memory

The memory assigned to every VM in the infrastructure is configured as Static, to avoid any possible occurrence of VMs competing for the same resources.

Hyper-V configuration of NUMA

Non-Uniform Memory Access (NUMA) capabilities were enabled on the PowerEdge R720 physical hosts (Node Interleaving disabled in the server BIOS) to allow memory access across CPUs. These R720s have two NUMA nodes each managing 192GB of memory. The Hyper-V NUMA Spanning setting is left enabled (by default).

Guest VMs disks

The virtual disks in use to host the operating system of each VM are VHDX fixed type disks. The VHDX files for the VMs hosted from the R620 hypervisor are deployed on the second pair of local RAID-1 disks, all the VMs hosted by the R720 hypervisors are deployed on the SAN as described below:

- VHDXs for the system disks deployed on the PS6110XV array
- VHDXs for the Exchange data deployed on the PS6210E arrays (host initiator scenarios)

Host network adapters and virtual network configuration

The host network adapters providing connectivity for the hosts and the VMs are configured as listed.

- Two physical network adapters, sourced from the onboard Broadcom 1GbE ports, providing connectivity for host domain access and management
- Two physical network adapters, sourced from the onboard Broadcom 1GbE ports, connected independently via two virtual network switches, and providing connectivity for the VMs to both domain and intra-VM traffic
- Two (on the R620 or the R720s) physical network adapters, sourced from the Broadcom 10GbE ports, connected independently via two virtual network switches, and providing access to the iSCSI SAN from the guest environments for these use cases configuration.
- Two (on the R720s) physical network adapters, sourced from the Broadcom 10GbE ports, connected independently, and providing access to the iSCSI SAN from the host for these use cases configuration.
- MPIO (Multi-path I/O) enabled and provided by EqualLogic DSM module, either for the 10GbE ports configured for host or guest initiator iSCSI traffic

The non-default settings implemented on all physical network adapters include:

- Jumbo frames enabled
- Flow Control enabled
- Large send offload enabled
- Receive and transmit buffers maximized
- Virtual Machine Queues (for each network adapter used by a virtual network switch)
- SR-IOV enabled (only for the 10GbE ports used for guest iSCSI traffic)

The non-default settings implemented on the virtual network switches used for iSCSI traffic include:

- Single-Root I/O Virtualization (SR-IOV) enabled

Virtual network adapter configuration

The assignment of the virtual network adapters of the VMs is configured as listed below:

- Two virtual network adapters to access the LAN traffic with switch independent mode NIC teaming by LBFO (All active adapters)
- Two virtual network adapters to access the iSCSI traffic independently bound to the corresponding virtual network switch

The hardware acceleration settings implemented on the virtual network adapters are:

- Virtual machine queue enabled on every adapter
- Single-Root I/O Virtualization (SR-IOV) enabled only for the adapters used for iSCSI traffic

The non-default settings implemented on the virtual network adapters used for iSCSI traffic include:

- Jumbo frames enabled
- Large send offload enabled
- Receive and transmit buffers maximized

B   Microsoft Jetstress considerations

Microsoft Exchange Server Jetstress 2013 is a simulation tool able to reproduce the database and logs I/O workload of an Exchange mailbox role server. It is usually used to verify and validate the conformity of a storage subsystem solution before the full Exchange software stack is deployed. Some elements worth being considered about Microsoft Jetstress are:

- Does not require and should not be hosted on a server where Exchange Server is running
- Performs only Exchange storage access and not host processes simulations. It does not contribute in assessing or sizing the Exchange memory and processes footprints
- Is an ESE application requiring access to the ESE dynamic link libraries to perform database access. It takes advantage of the same API used by the Exchange Server application software stack and as such it is a reliable simulation application
- Runs on a single server. When a multiple servers simulation is required, the orchestration of the distributed instances has to be fostered by external management tools
- Requires, and provides, an initialization step to create and populate the database/s that will be used for the subsequent test phases. The database/s should be planned of the same capacity as the one/s planned for the Exchange Server future deployment
- Its topology layout includes number and size of simulated mailboxes, number and placement of databases and log files, number of database replica copies (it simulates only active databases)
- While carrying out a mailbox profile test, it executes a pre-defined mix of insert, delete, replace and commit operations against the database objects during the transactional step, then it performs a full database checksum
- Collects Application and System Event Logs, performance counter values for the criteria metrics of both operating system resources and ESE instances during transactional and DB checksum phases. It then generates a detailed HTML-based report
- Throttles the disk I/O generation using the assigned IOPS per mailbox, thread count (global per all databases) and SluggishSessions threads property (fine tuning for threads execution pace)

C   Additional resources

Support.dell.com is focused on meeting your needs with proven services and support.

DellTechCenter.com is an IT Community where you can connect with Dell Customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Configuration Guide:
  http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19852516/download.aspx

- Sizing and Best Practices for Deploying Microsoft Exchange Server 2013 on VMware vSphere and Dell EqualLogic PS6110E Arrays
  http://en.community.dell.com/techcenter/extras/m/white_papers/20438750/download.aspx

- Enhancing Microsoft Exchange Server 2010 Data Protection and Availability using Dell EqualLogic Snapshots
  http://en.community.dell.com/dell-groups/dtcmedia/m/mediagallery/19961113/download.aspx

Referenced or recommended Microsoft publications:

- Tools for performance and Scalability Evaluation:

  http://technet.microsoft.com/en-us/library/dd335108.aspx

- Exchange 2013 Storage Configuration Options:

  http://technet.microsoft.com/en-us/library/ee832792%28v=exchg.150%29.aspx