Microsoft Exchange 2010 on a Hyper-V Virtual Infrastructure Supported by Dell EqualLogic SANs

A Dell EqualLogic Best Practices Technical White Paper

Dell Storage Engineering
February 2013

Table of contents

Acknowledgements

This best practice white paper was produced by the following members of the Dell Storage team:

Engineering: Danilo Feroce

Technical Marketing: Jim Salvadore

Editing: Margaret Boeneke

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

Virtualization technologies are an optimal answer to address the need to reduce IT operational costs. These technologies were originally introduced with the aim to consolidate multiple functional services into a single physical system and to increase hardware utilization. In addition to their basic features, with additional maturity and growth these platforms now target:

- Enhancing application and services availability
- Agility and simplification of managing complex infrastructures
- Streamlining provisioning of additional resources in rapidly changing environments
- Providing differentiated tactics for disaster recovery and business continuity

This changing ecosystem requires substantial scrutiny of the workload in a virtual infrastructure to maintain, or widen, the effectiveness of the application when it comes to network or storage planning. A messaging infrastructure built on Microsoft® Exchange has become one of the most critical corporate-wide services. The 2010 version provides a range of improved capabilities and also the ability to dramatically expand the mailbox storage repository for end-users via the redesign of its storage access mechanisms.

This paper answers two important questions: How can the deployment outcomes of the new Exchange features be planned during the design of a storage subsystem in a virtual infrastructure? How can a system take advantage of the updated Dell™ EqualLogic™ PS6100 Series arrays with the new Exchange 2010 workload?

The deployment variables of the Exchange 2010 mailbox server role were examined and their effect on the Exchange main performance indicators was evaluated. Some decisive findings of this research are:

- The use of a host iSCSI initiator causes a small efficiency penalty because each storage I/O is required to cross a greater number of logical layers before reaching the SAN, as opposed to a guest iSCSI initiator.
- RAID policy analysis shows a small difference in performance in favor of RAID 50 versus RAID 6 during Exchange workloads. The pros and cons evaluation of this type of difference should be weighed against characteristics such as tolerance to disk failure or available capacity.
- The distribution of user mailboxes with a predefined workload across fewer databases creates the benefit of a lower overall amount of IOPS performed by the mailbox server. Nevertheless, the concentration of many users in a single database should be carefully weighed against the administrative tactics and the data protection measures.
- Each of the PS6100XV 3.5", PS6100X, and PS6100E models offer a specific combination of disk performance and available capacity. The selected array should be fitted to the individual workload.
- Scaling the effectiveness of an EqualLogic SAN either up (increasing the user workload) or out (increasing the number of arrays) always provides linear growth when properly planned.
- The number of DAG nodes and database copies greatly affects the additional workload enforced during operation activities like seeding or re-seeding of databases.

1 Introduction

The EqualLogic PS6100 Series arrays add a heightened capacity and increased overall performance while preserving the same peer storage architecture, the ease of use and management, the broad out-of-the-box feature set, and software portfolio as its predecessors. These include scalability without fork-lift upgrades, native load balancing, point-in-time snapshots, integrated replication, and application layer software integration and monitoring. The pronounced redesign of Microsoft Exchange 2010 and its storage functionalities have provided organizations with the opportunity to promote mailboxes with expanded storage capacity while retaining or improving the previous efficiency and capabilities.

The benefits of a lower Total Cost of Ownership (TCO) with virtualization technologies and private clouds provides a compelling push for IT departments to innovate their current infrastructures. IT professionals look forward to the right balance of high availability (HA), reliability and scalability of their virtualized solutions.

The advantages of these solutions built within a smaller physical footprint, with more efficient and more granularly managed power consumption and more flexible management under a single pane of glass, encourage the shift towards consolidated infrastructures powered by blade servers and provisioned by virtualization technologies.

1.1 Purpose and scope

This white paper describes the results of a study performed to provide guidance and best practices for deploying Microsoft Exchange Server 2010 on a virtual infrastructure powered by Microsoft Hyper-V technology and built on Dell PowerEdge™ blade servers and an EqualLogic SAN. It helps Exchange and SAN administrators to understand their messaging workload and predict their SAN utilization. The scope of this paper is restricted to a local datacenter topology and does not include server sizing exercises.

1.2 Target audience

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, and/or implementing Microsoft Exchange Server infrastructures and who would like to investigate the benefits of using EqualLogic storage. This document assumes the reader is familiar with Microsoft Exchange functions, EqualLogic SAN operation, and Microsoft Hyper-V architecture and system administration.

1.3 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 and contributes its storage space to the entire 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.

Parent and Child Partitions: Identifies the logical units of isolation supported by the Hyper-V hypervisor. The parent, or root, partition hosts the hypervisor itself and spawns the child partitions where the guest virtual machines reside.

Virtual Machine (VM): An operating system implemented on a software representation of hardware resources (processor, memory, storage, network, etc.). 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.

Synthetic drivers: Supported with the Hyper-V technology, these drivers leverage more efficient communications and data transfer between the virtual and physical hardware, as opposed to legacy emulated drivers. Synthetic drivers are only supported in newer operating system versions and allow the guest VMs to become enlightened, or aware, that they run in a virtual environment.

Virtual network: A network consisting of virtual links as opposed to wired or wireless connections between computing devices. A virtual network is a software implementation similar to a physical switch, but with different limitations. Microsoft Hyper-V technology implements three types of virtual networks allowing the connectivity between VMs and devices external to the root host (external), the connectivity between the VMs and the host only (internal), or the VMs running on a specific host only (private).

VHD: File format for a Virtual Hard Disk in a Windows Hyper-V hypervisor environment.

Non-Uniform Memory Access (NUMA): A multiprocessing architecture in which memory is associated to each CPU (local memory, accessed faster than memory of other processor) as opposed to SMP where memory is shared between CPUs.

Microsoft Exchange Database Availability Group (DAG): A pool of networked Exchange mailbox servers that hosts multiple copies of the same Exchange databases.

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.

Process: An instance of a computer program or application that is being executed. It owns a set of private resources: image or code, memory, handles, security attributes, states, and threads.

Thread: A separate line of execution inside a process with access to the data and resources of the parent process. It is also the smallest unit of instructions executable by an operating system scheduler.

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

2 Microsoft Exchange and the storage subsystem

Microsoft Exchange Server has a diversified set of components and services that cooperate to support the disparate requirements to design and deploy a messaging infrastructure within an organization. The mailbox server role in an Exchange infrastructure has the most impact on storage resources because it ultimately governs the storage, retrieval, and availability of user data to be routed and presented to the rest of the infrastructure.

Appropriate sizing of the mailbox role servers is the primary best practice that can prevent performance bottlenecks and avoid the administrative overhead required to redesign the deployment topology to adapt to new or changed user requirements. To understand the interaction between the Exchange mailbox server role and the storage subsystem, the underlying logical components in the hierarchy of Microsoft Exchange Server 2010 should be examined.

2.1 Exchange store elements

The access to mailbox databases (or public folder databases, when implemented) is the primary element that generates I/O on a storage subsystem. However, while a database is a logical representation of a collection of user mailboxes, it is also an aggregation of files on the disk which are accessed and manipulated by a set of Exchange services (for example, Exchange Information Store, Exchange Search Indexer, Exchange Replication Service, and Microsoft Exchange server) following a different set of rules.

Database file (*.edb): The container for user mailbox data. Its content, broken into database pages of 32 KB, is primarily read and written in a random fashion 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 2010 is 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 (create, modify, delete messages, and others) 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 instead read solely when in a replicated database configuration within a DAG or in the event of a recovery.

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.

Microsoft Exchange Server uses a proprietary format called Extensible Storage Engine (ESE) to access, manipulate, and save data to its own mailbox databases. The same format is now 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 releases 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.

For additional information about the Exchange 2010 Store, refer to Microsoft documentation *Understanding the Exchange 2010 Store*, available at: http://technet.microsoft.com/en-us/library/bb331958.aspx

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 database maintenance, which includes dumpster 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 Database Availability Group configuration.

2.2 Defining the configuration and deployment variables

After determining the I/O activities generated against the storage subsystem for a single database, the factors that influence the overall I/O footprint of an entire Exchange mailbox server role and the variables of a deployment should be taken into consideration.

Mailbox usage profiles: Denote the usage characteristics of a mailbox (for example: send, receive, open, or delete items). They are commonly defined by the amount of messages sent and received per day and the average message size. It can also be explained in terms of transactional IOPS per mailbox when considering the size of database cache allocated per mailbox.

Mailbox size: Defines the maximum size a mailbox is allowed to grow and can be enforced by a quota policy. In more general terms, it is the average mailbox size in a corporate messaging environment. It primarily affects the capacity requirement of a mailbox role server and the planned size of the databases hosted by the server. Moreover, it influences the IOPS response due to the amount of physical disk surface that must be accessed to retrieve or store the data.

Number of databases: Designates the database layout and mailbox distribution across the databases. Each Exchange database is managed as a single administrative unit and is assisted by a set of services with a 1:1 ratio (defragmentation, maintenance, logs generation).

Database and Log placement: Indicates whether the *.edb and log files reside on the same volume or are deployed on isolated volumes. There are two historical reasons for the requirement to split database and logs into different volumes (or physical drives):

- Performance: The different I/O pattern of these two streams of data (random reads/writes versus sequential writes) and the aim to associate them with the most fitting storage device (for example because of rotational speed or RAID policy).
- Reliability: The simultaneous loss of both the database and logs could jeopardize the recoverability of user data depending on the data protection tactic in place.

Now, the Exchange Server 2010 I/O footprint reduction, when compared with former versions, lessens the requirement to split the database and log placement. In addition, the use of a SAN as a storage subsystem provides different approaches such as EqualLogic Smart Copies to the data protection of the Exchange mailbox server role that do not require the volume split mentioned above. For additional references refer to the white paper mentioned in the notes box at the end of this section.

For additional information about Exchange data protection options with Dell EqualLogic SAN refer to the white paper *Best Practices for Enhancing Microsoft Exchange Server 2010 Data Protection and Availability using Dell EqualLogic Snapshots*, available at: http://en.community.dell.com/techcenter/storage/w/wiki/2633.enhancing-microsoft-exchange-server-2010-data-protection-and-availability-with-equallogic-snapshots-by-sis.aspx

Mailbox count: Represents the number of user mailboxes hosted by the mailbox role server. When the mailbox count increases, the amount of IOPS, capacity allocated, and logs generated increase. Furthermore, when a large number of mailboxes are hosted by a single server, this intensifies the demand for a highly available solution due to the risk of widely distributed loss of service in a messaging organization.

High availability footprint: Refers to the overhead of having a highly available solution in place with data replication involved. See Section 2.3.

Data protection footprint: Identifies the planned amount of IOPS used by the solution that protects the mailbox database data. Since a customized solution is usually tailored to each distinct production environment (for example 24x7, or 9 to 5) the impact is greatly variable and depends upon the degree of parallelism between the regular data access and the data protection access.

2.3 Considerations for Exchange DAG

A 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. When DAG is implemented, it creates some deviations in the storage access patterns and in the Exchange memory cache behavior.

Transaction Logs access when using DAG configuration adds to the conventional sequential write pattern a sequential read access pattern required to perform the replication activities.

Log Checkpoint depth refers to the amount of logs written to the disk and that contain transactions not yet flushed to the database file. It is usually set at 20 in a standalone server installation and is increased to 100 when implementing a DAG configuration. This change reduces the write I/O for the given database because user activity changes are combined in memory (coalescing), and thus the overall I/O is reduced.

3 Test topology and architecture overview

This paper presents a set of findings from tests conducted on a Microsoft Windows infrastructure and an Exchange organization built on Windows Server with Hyper-V technology accessing storage resources provided by an EqualLogic SAN. Multiple simulation tools, both Microsoft Jetstress and Loadgen, were used to generate the workloads required for this study.

The architecture deployed consists of two different sets of components residing within the same topology and able to support either the servers running the Jetstress simulated mailbox servers or the complete messaging services required to run a medium-sized organization (up to 5,000 users).

3.1 Functional system design

The functional elements for the infrastructure supporting the Jetstress-based tests are shown in Figure 1.

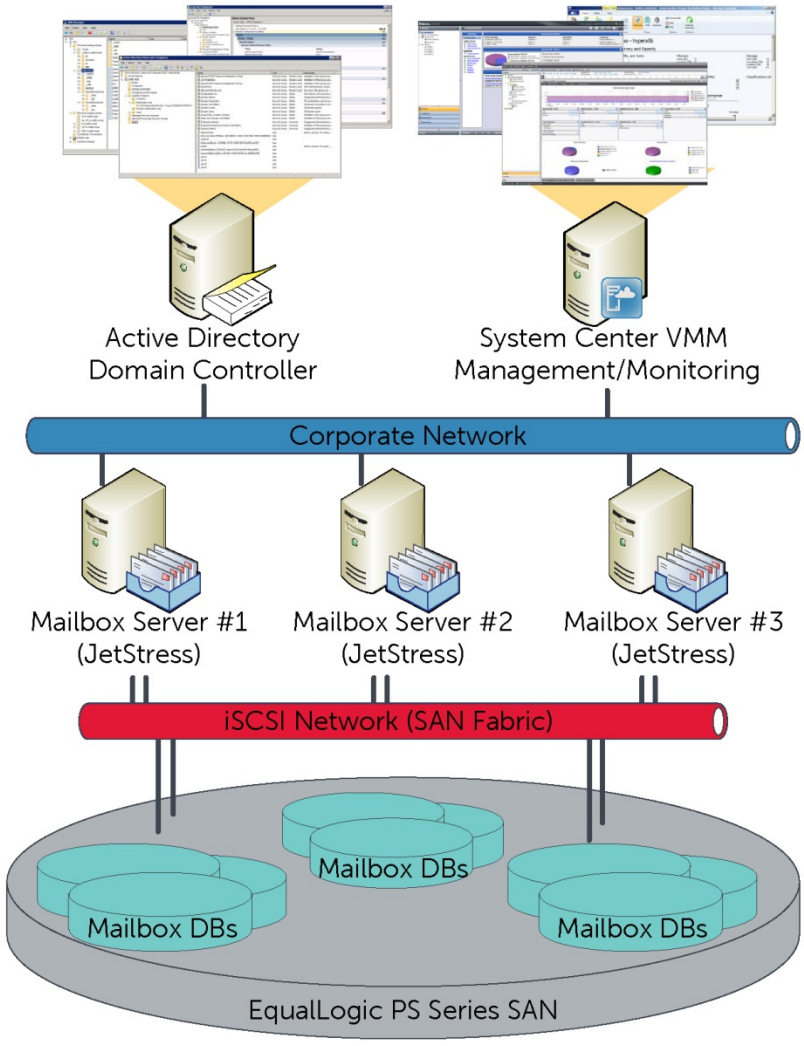

Figure 1 Functional system design for Jetstress tests

The main elements of the design are:

- Single Active Directory forest, single domain, single site (not strictly required for the tests)
- Centralized management and monitoring with dedicated resources
- Building block design approach for mailbox role servers with Jetstress
- Separated network design for traffic isolation between LAN and iSCSI

The architecture has then been modified to include the additional elements required to support the complexity of an Exchange organization deployed with highly available building blocks, as shown in Figure 2.

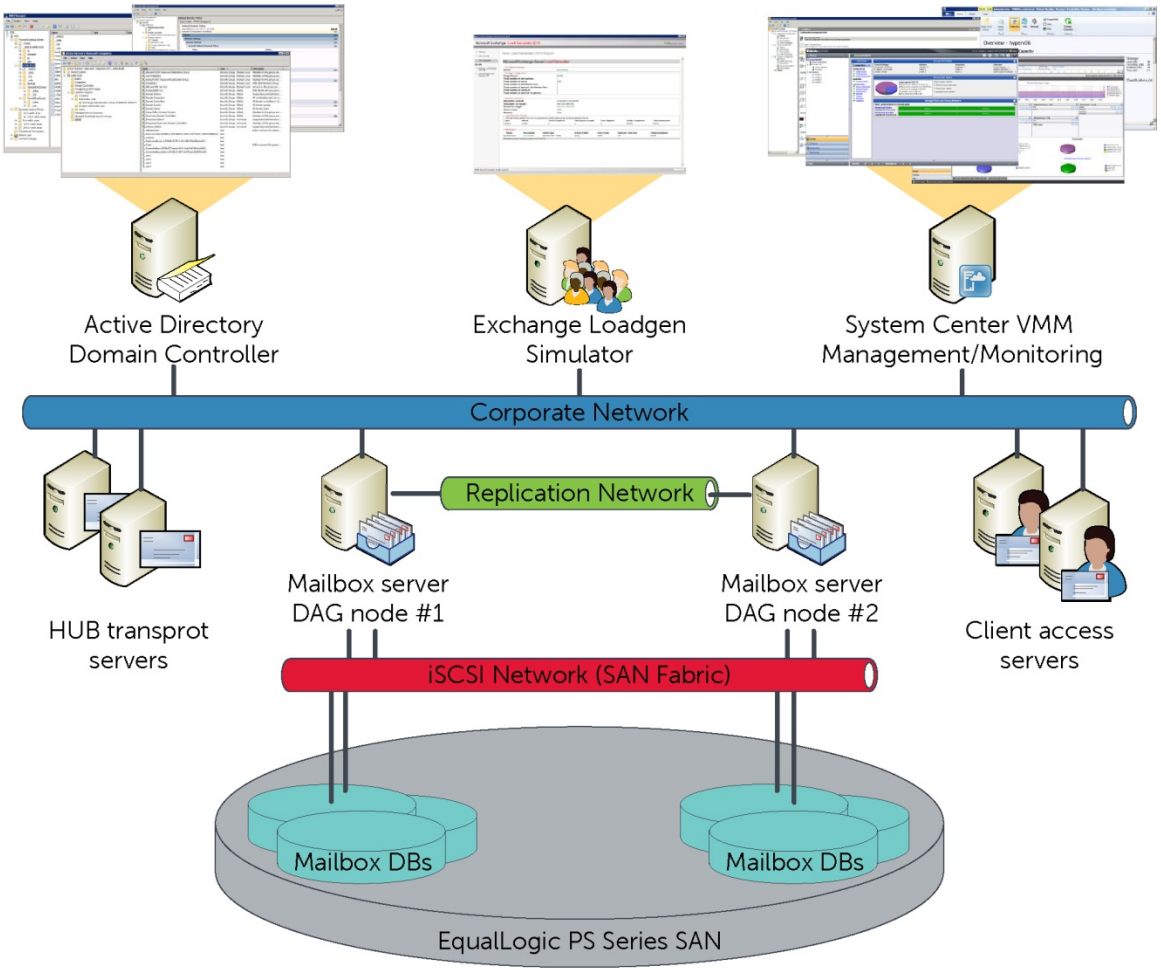

Figure 2 Functional system design for Exchange Server tests

Additional components include:

- Redundant HUB transport server services and resilient Client Access servers (load balanced)
- Mailbox database servers participating in a DAG for HA and resiliency
- Dedicated resources for client-side simulation with the Exchange Loadgen tool

3.2 Physical system configuration

The physical components of the test infrastructure were configured as show in Figure 3.

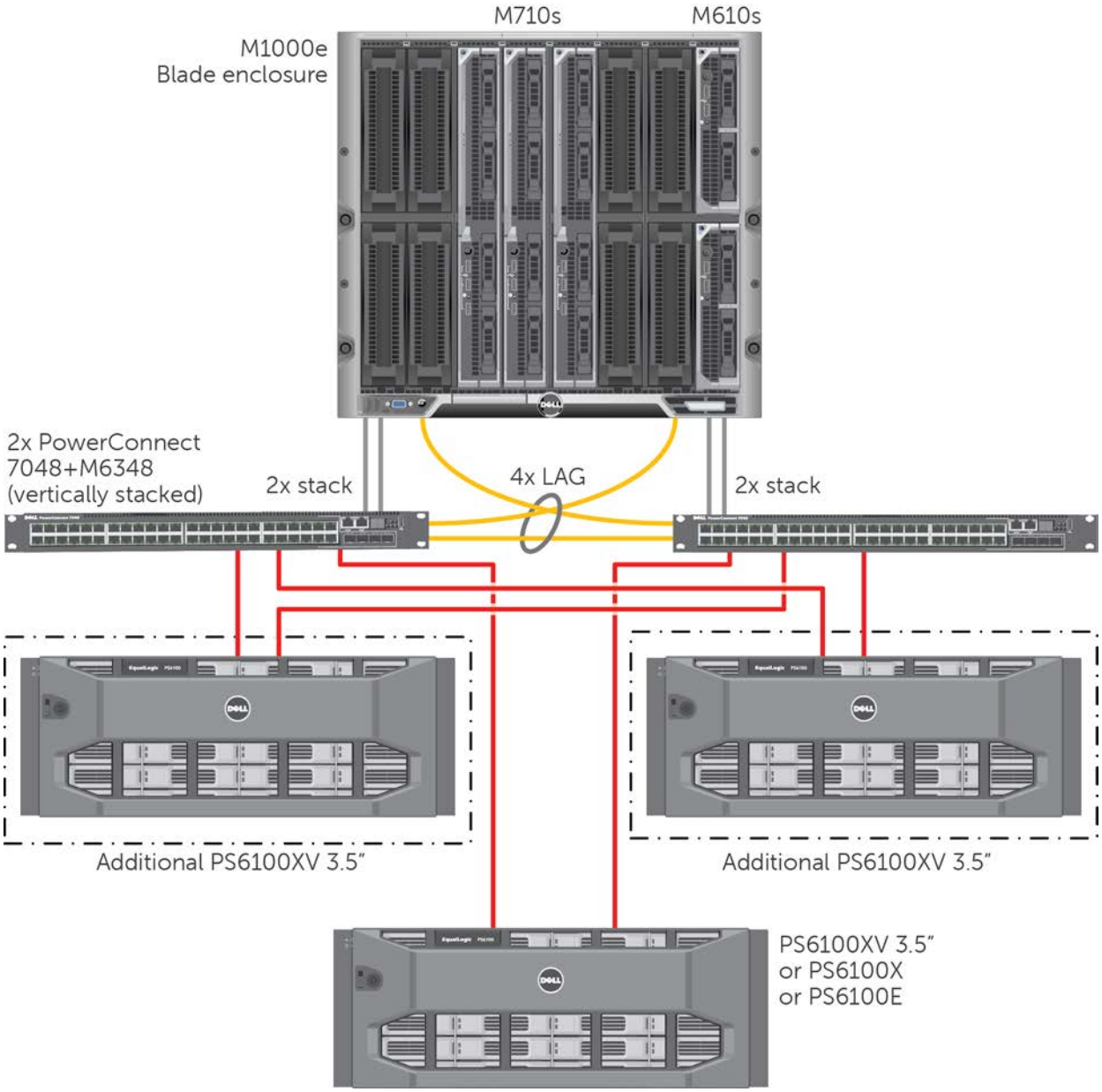

Figure 3 Physical system design

The solution is deployed on Dell Blade servers with the aim of a greater datacenter density and flexibility. The main components of the physical deployment are:

- Single M1000E Blade enclosure with redundant chassis management controllers (CMC) and fully populated power supplies configured for redundancy

- Single EqualLogic iSCSI SAN provisioned with one of the following array models (exchanged for each test): PS6100 3.5", PS6100X, or PS6100E
- Single EqualLogic iSCSI SAN with up to three PS6100XV 3.5" arrays for the test to scale out the SAN to multiple arrays or to support multiple data copies of the DAG nodes
- Dual PowerConnect M6220 Ethernet switches (stacked) to support LAN IP traffic (Fabric A)
- Dual PowerConnect M6348 Ethernet switches (vertically stacked in pairs with the PC7048R switches) to support the iSCSI data storage traffic on the server side (Fabric B)
- Dual PowerConnect 7048R Ethernet switches (vertically stacked in pairs with the M6348) to support the iSCSI data storage traffic on the SAN side
- Link Aggregation Group (LAG) consisting of four fibre connections (two from each switch) between the Fabric B M6348s and the top of rack (ToR) PC7048R switches in a mesh pattern

Note: More details of the test configuration, including a hardware and software list, SAN array characteristics, hypervisor and VMs relationship, network connections, and blade switch fabric paths are provided in Appendix A.

3.3 Storage layout

The EqualLogic SAN arrays and the volumes underlying the Exchange databases are configured with:

- One EqualLogic group configured with one unit of each array model (exchanged for each test): PS6100 3.5", PS6100X, or PS6100E.
- One EqualLogic group configured with one to three array members for the simulation to scale the SAN to multiple arrays.
- One storage pool defined within the group that includes the single member or all the members of the group for the test case of multiple arrays.
- The RAID policy specified as part of each test case.
- Five volumes created in the pool, unless specified in the test case. One volume dedicated to each Exchange mailbox database and a set of log files with a 1:1 ratio.

Figure 4 reflects the common volume layout implemented on the EqualLogic SAN.

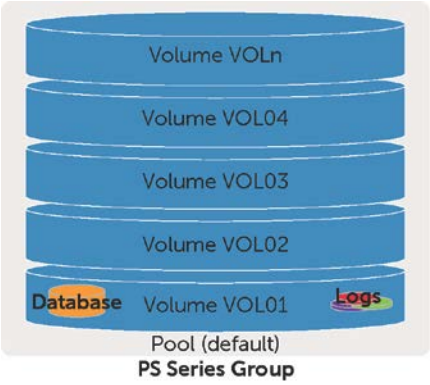

Figure 4 Storage layout

4 Trend of Exchange deployment variables

The performance trends of an EqualLogic SAN and of the Hyper-V virtual infrastructure were evaluated under the load of a simulated Exchange mailbox role server. The following sections summarize each of the variations tested to understand these trends.

To forecast the storage requirements of a present or future deployment, begin with the analysis of the topics listed below:

- Define an average usage profile for a mailbox user in the organization
- Define the mailbox quota or cap the organization plans to enforce
- Average mailbox count per database planned for the deployment
- HA options to be considered for the mailbox service (DAG) and their workload overhead
- Collocation of mailbox database files on the SAN and methodology to access them (initiators)
- Disks failure tolerance, available capacity, and performance of the RAID policy selected
- Storage capacity and performance of the array model selected against the Exchange type workload
- Number of users per deployment building block and space for future growth
- Horizontal scalability of the SAN and the virtual infrastructure

The reference Exchange mailbox server role configuration used is detailed in Table 1. For each test completed, a description of the variations applied to this reference baseline is provided in each section.

Table 1 Reference configuration for Microsoft Jetstress 2010 tests

Reference configuration: Test variables under study	
Number of simulated mailboxes/users	5000 concurrent users
Number of databases	5 databases (active)
Mailbox allocation	1000 mailboxes per each mailbox database
Database size	1TB each
RAID policy	RAID 6
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
Number of units, Array model, SAN configuration	1x PS6100XV 3.5", one single pool (default)
Reference configuration: Consistent factors across each test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Mailbox size	1GB each
Number of database replica copies	2 (2 node DAG)
Background database maintenance	Enabled
Windows Disk/Partition, File System	Basic disk, GPT partition, default alignment NTFS, 64KB allocation unit size
Test duration	2 hours + time required to end DB checksum (for Jetstress based tests)

Below is a list of metrics and pass/fail criteria recorded while completing the tests. Most of this information is outlined by the Jetstress tool or is verified through Windows Performance Counters and Dell EqualLogic SAN Headquarters while a Jetstress or Loadgen simulation is running. Microsoft data around thresholds for storage validation are reported as well.

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

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

Logs Writes Average Latency (msec) is the average length in time to wait for a log file write operation (sequential writes). It should be less than 10 milliseconds 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 tracked separately instead.

Differential IOPS are the IOPS generated for the DB maintenance 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 the three elements above: achieved transactional IOPS, Logs IOPS and differential 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 tools, Microsoft Jetstress and Loadgen 2010, refer to Appendix B.

4.1 Characterize the impact of the iSCSI software initiator collocation

The goal of the initial iSCSI initiator type analysis was to establish the Exchange KPI trend, IOPS ratios, and their relationship when using two different collocations for the iSCSI software initiator enabling the access to the SAN volumes, while maintaining the other factors. Table 2 shows the configuration parameters for this test.

Table 2 Test parameters: iSCSI software initiator collocation

Reference configuration: Test variables under study	
iSCSI initiator software collocation	Host initiator (residing in the Hyper-V root) or Guest initiator (residing in the VMs)
Reference configuration: Consistent factors across this test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Number of simulated mailboxes/users	5,000 concurrent users
Number of databases	5 databases (active)
Mailbox allocation	1,000 mailboxes per each mailbox database
Mailbox size	1GB each
Database size	1TB each
Number of database replica copies	2 (2 node DAG)
RAID policy	RAID 6
Number of units, Array model, SAN configuration	1x PS6100XV 3.5”, one single pool (default)

In a typical iSCSI based SAN, the initiator is regarded as the client which is accessing the storage resources located on an iSCSI server, or target, by sending SCSI commands over an IP network. The initiator falls into two broad types: software initiator (software implementation installed within the operating system running the initiator) or hardware initiator (a dedicated hardware resource, most commonly a host bus adapter).

In the infrastructure built for our tests we used software initiators only, provided natively by Windows Server 2008 R2 and integrated by the EqualLogic Host Integration Tool Kit. The two configurations validated are based on a different collocation of the software initiator.

1. Guest initiator: software located and running on the guest VMs, which allows you to directly connect to the volumes residing on the EqualLogic SAN from the virtual network adapters of the guest. The settings of the VMs include four additional virtual network adapters dedicated to SAN traffic. The host hypervisor is not aware of the type of traffic traversing the VMBus adapters.
2. Host initiator: software located and running on the host hypervisor, which allows you to connect to the volumes residing on the EqualLogic SAN from the root partition of Hyper-V through four physical network adapters dedicated to the SAN traffic. VHD fixed-sized files are created on each of the SAN volumes attached to the host and then added as SCSI disks to the settings of the VMs. The VMs are not aware of where their disks reside, either on local storage or the SAN.

Figure 5 shows the results collected from the Exchange Jetstress simulations of the two different iSCSI initiator collocations.

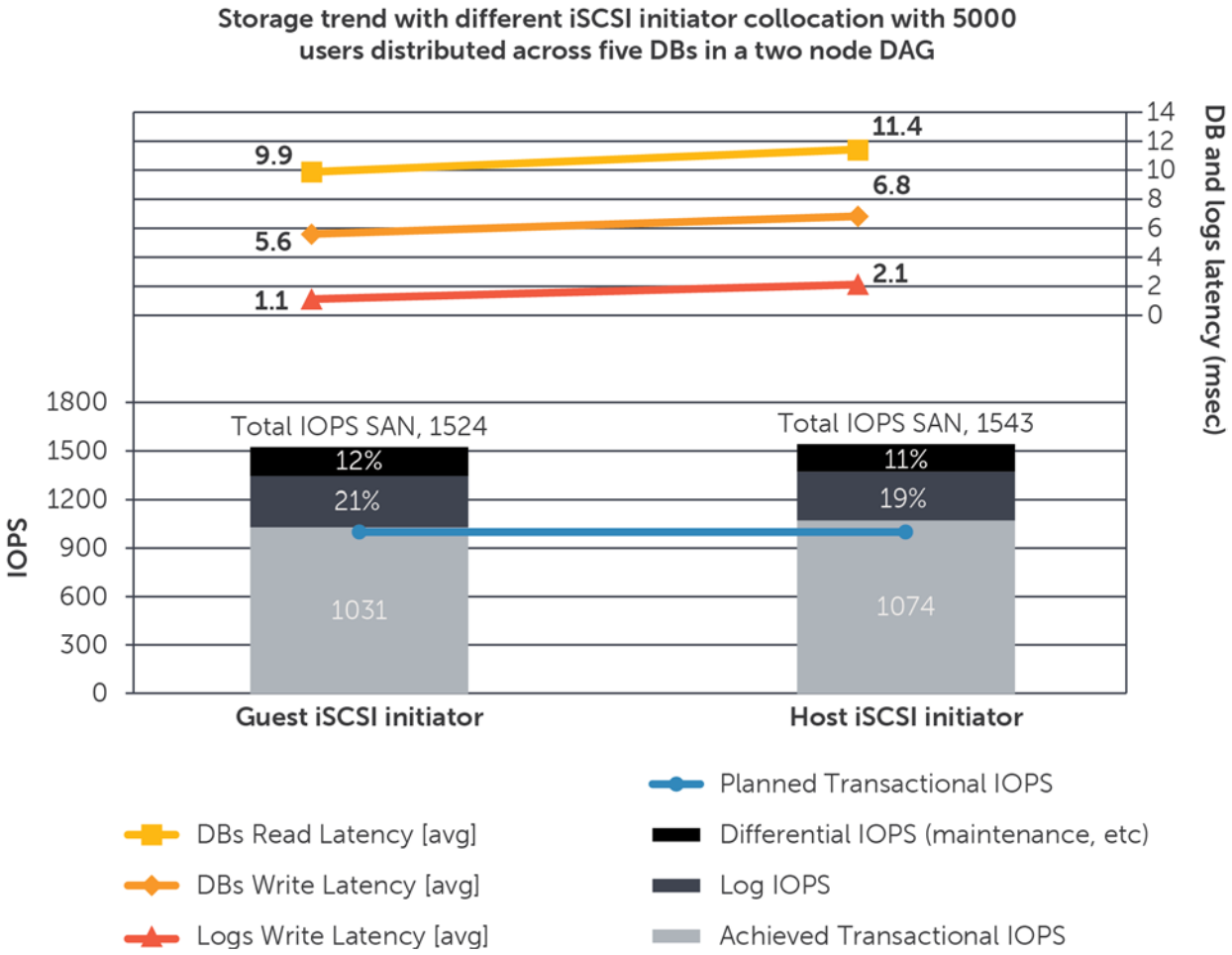

Figure 5 Storage trend with different iSCSI software initiator collocation

Note: The Jetstress tool throttles its storage request activities during the simulations by way of two tuning parameters (threads and sluggish sessions). The planned IOPS and the achieved IOPS cannot exactly match due to the variance of the Jetstress threads efficiency during each iteration of the simulations, regardless the effort to configure them to have a perfect match. As a result the latencies recorded must be regarded as the criteria to measure the achieved and not the planned IOPS. To have a perfect comparison of these latency values, they should be normalized from achieved towards planned IOPS.

Table 3 reports the results recorded during the iSCSI initiators test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed and not by the transactional load alone.

Table 3 Test results: iSCSI software initiators collocation with KPI increase or decrease relationship

	Guest iSCSI initiator	Host iSCSI initiator
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [+3%]	1074 IOPS [+7%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [+52%]	1543 IOPS [+54%]
DBs Read Latency [average]	9.9 msec	11.4 msec
DBs Write Latency [average]	5.6 msec	6.8 msec
LOGs Write Latency [average]	1.1 msec	2.1 msec
Resulting latencies after normalization to the Planned Transactional IOPS of 1000 IOPS		
DBs Read Latency normalized	9.6 msec [acting as baseline]	10.6 msec [+11%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	6.3 msec [+17%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	2.0 msec [+84%]

The outcomes show a small efficiency penalty at the host iSCSI initiator level, consistent with the obligation of each storage access operation to cross a greater number of logical layers before reaching the SAN. The greatest decline was reported by the write activities and can be attributed to the nature of the log file access which is based on small size operations.

4.2 Characterize the RAID policy in use

The goal of the RAID policy analysis was to establish the Exchange KPI trend, IOPS ratios, and their relationship when implementing different RAID policies on the EqualLogic SAN array, while maintaining the other factors. Table 4 shows the configuration parameters for this test.

Table 4 Test parameters: RAID policies variation

Reference configuration: Test variables under study	
RAID policy	RAID 6 / RAID 50
Reference configuration: Consistent factors across this test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Number of simulated mailboxes/users	5,000 concurrent users
Number of databases	5 databases (active)
Mailbox allocation	1,000 mailboxes per each mailbox database
Mailbox size	1GB each
Database size	1TB each
Number of database replica copies	2 (2 node DAG)
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
Number of units, array model, SAN configuration	1x PS6100XV 3.5", one single pool (default)

The EqualLogic PS Series 6100 arrays provide a range of different RAID policies and the ability to use spare drives to protect data. Each RAID level offers a distinct set of performance and availability characteristics dependent on the nature of the RAID policy and the workload applied.

- RAID 6, one or more dual parity sets
  The highest availability at the expense of random writes performance and rebuild impact.
  A heavy impact for RAID reconstruction in the case of drive failure.
  A total of 23 available drives on PS6100, with one spare drive.
- RAID 50, or striping over multiple distributed parity sets (RAID 5)
  A balance between performance and capacity.
  A moderate impact for RAID reconstruction in case of drive failure.
  A total of 22 available drives on PS6100, with two spare drives.
- RAID 10, or striping over multiple mirrored sets (RAID 1)
  The best performance for random access at the expense of an average capacity available.
  A minimal impact for RAID reconstruction in case of drive failure.
  A total of 22 available drives on PS6100, with two spare drives.
- RAID 5, while still supported, is not advised for critical data due to the limited level of resiliency.

Note: RAID implementations deprived of spare drives for protection, even if allowed on the EqualLogic PS Series, are not considered here and in the rest of this study.

EqualLogic PS arrays also support RAID policy conversion, which allows switching from one RAID level to another without disruption of the data present on the volumes, although the performance during the conversion might be degraded. The following rules apply for RAID conversion:

- Conversion from RAID 10 to RAID 50 or RAID 6 is supported
- Conversion from RAID 50 to RAID 6 is supported
- Conversion from RAID 6 or RAID 5 is not supported, a reset and initialization of the array is required

The two RAID policies selected for the test, reported in the Table 4 above, use the PS6100XV 3.5" as the reference array. The array was reinitialized to the factory level for each RAID level run, the volumes were recreated, and the Exchange Jetstress databases were redeployed. Specifically for this set of tests the configuration of Exchange Jetstress, in term of threads and sluggish sessions, remains strictly the same across the runs because the aim is to capture the IOPS handled by the array consistently with the RAID policy implemented. The RAID 10 policy was not evaluated due to the excessively lower available capacity when compared to other RAID policies and based on the demanding capacity requirements of Exchange Server 2010.

Figure 6 shows the results collected from the Exchange Jetstress simulations with different RAID policies implemented on the SAN.

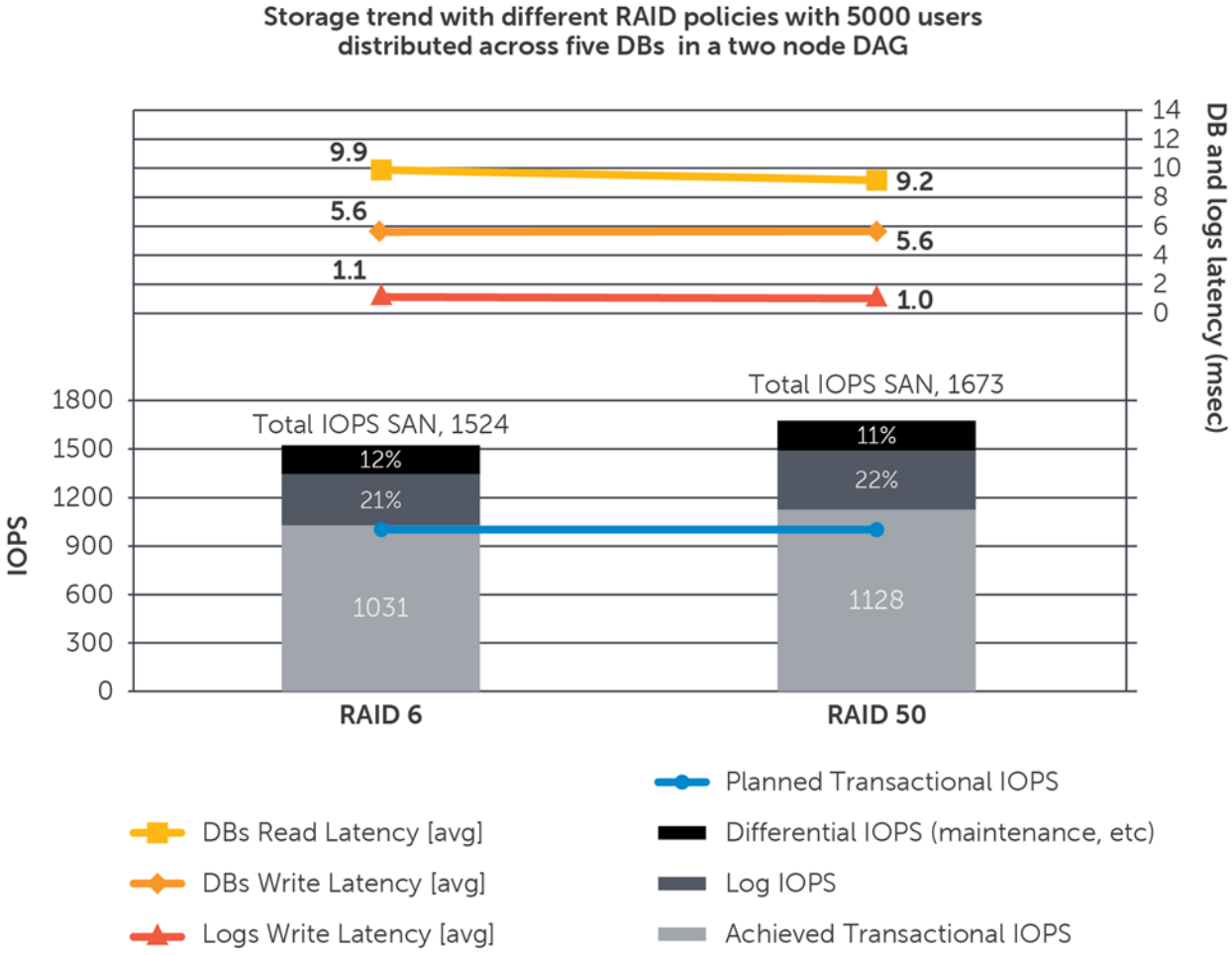

Figure 6 Storage trend with different RAID policies

Table 5 reports the numerical results recorded during the RAID policies test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 5 Test results: RAID policies variation with KPI increase or decrease relationship

	RAID 6	**RAID 50**
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [+3%]	1128 IOPS [+13%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [+52%]	1673 IOPS [+67%]
DBs Read Latency [average]	9.9 msec	9.2 msec
DBs Write Latency [average]	5.6 msec	5.6 msec
LOGs Write Latency [average]	1.1 msec	1.0 msec
Resulting latencies after normalization to the Planned Transactional IOPS of 1000 IOPS		
DBs Read Latency normalized	9.6 msec [acting as baseline]	8.1 msec [-15%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	5.0 msec [-8%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	0.9 msec [-17%]

The outcomes confirm the superior IOPS achievement of RAID 50 versus RAID 6, with a difference fluctuating between 8% and 17% depending on the type of access. The evaluation of this difference should be weighed against additional characteristics, such the tolerance to disks failure or the available capacity.

4.3 Database volumes layout

The goal of the database deployment layout analysis was to establish the Exchange KPI trend, IOPS ratios, and their relationship when changing the number of mailbox databases (keeping the SAN volumes / databases ratio 1:1), while maintaining the remaining factors. Table 6 shows the configuration parameters for this test.

Table 6 Test parameters: Database deployment layout

Reference configuration: Test variables under study	
Number of databases	5 / 10 / 20 databases (active)
Mailbox allocation	1,000 / 500 / 250 mailboxes per each mailbox database
Database size	1TB / 500GB / 250GB each
Reference configuration: Consistent factors across this test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Number of simulated mailboxes/users	5000 concurrent users
Mailbox size	1GB each
Number of database replica copies	2 (2 node DAG)
RAID policy	RAID 6
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
Number of units, Array model, SAN configuration	1x PS6100XV 3.5", one single pool (default)

Exchange Server provides the ability to implement multiple databases on the same server with a limit imposed by the licensed version of Exchange Server installed (a maximum of 5 databases for Exchange Standard edition and a maximum of 100 databases for Exchange Enterprise edition). The number of user mailboxes hosted in a single database is not bound by a declared limit. It is an informed decision of the Exchange administrators to select how many users per database are deployed. The online mailbox move feature allows administrators to seamlessly redistribute mailboxes across databases or mailbox servers as well as postpone or lift constraints imposed by decisions made early in the design and deployment stage of the messaging infrastructure.

The three deployment scenarios selected for the test (and listed in the Table 6 above) specify an increasing number of databases while maintaining the total amount of users per mailbox server at 5,000. The progressive changes in the configurations include number of databases, users per database, and the resulting database size shifting from five databases (1,000 users per DB, 1TB DB), ten databases (500 users per DB, 500GB DB), and then to 20 databases (250 users per DB, 250GB DB).

Figure 7 shows the results collected from the Exchange Jetstress simulations when changing the number of deployed mailbox databases.

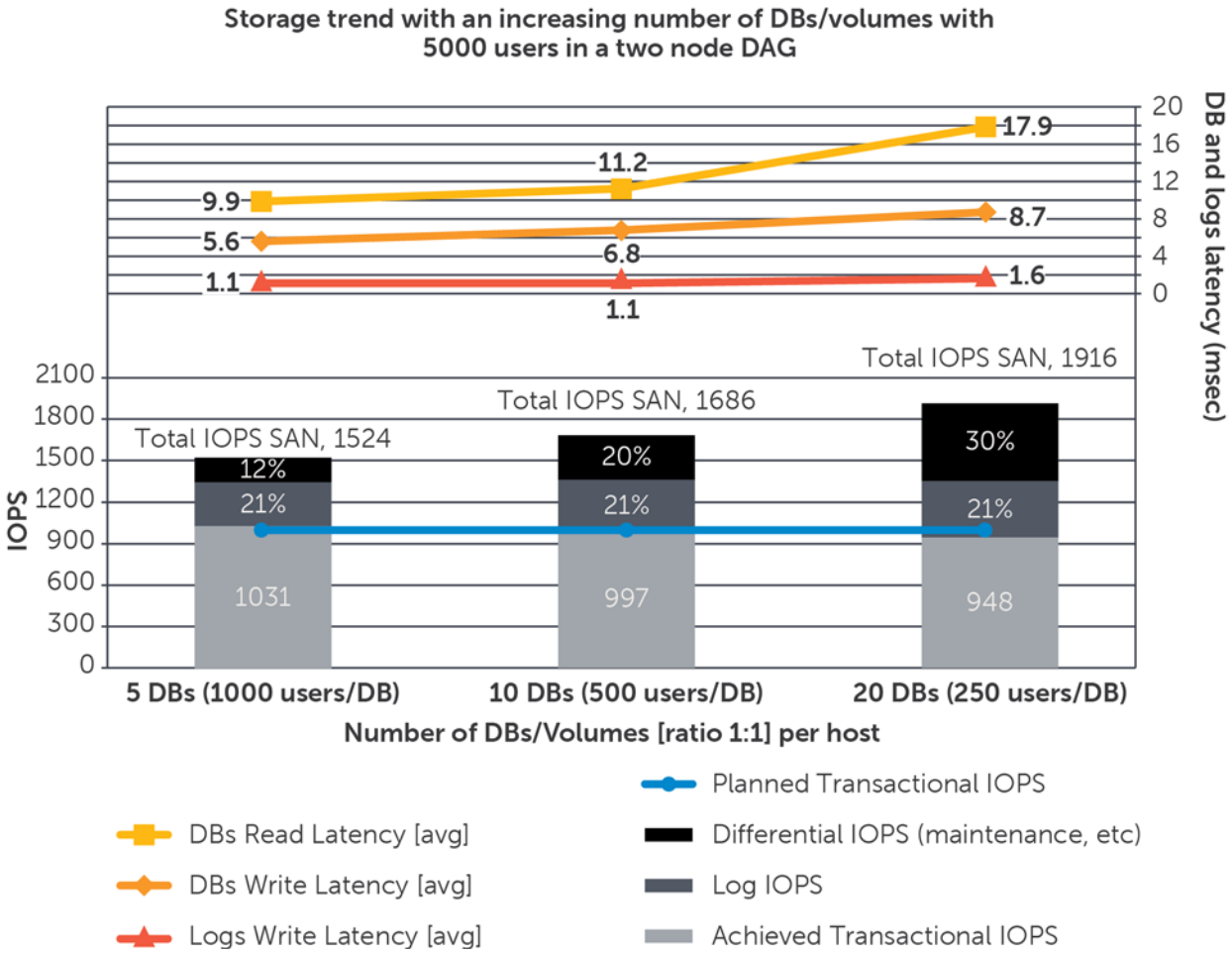

Figure 7 Storage trend with increasing number of DBs/Volumes with 5,000 users in a two node DAG

Table 7 reports the numerical results recorded during the increase of the number of DBs test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 7 Test results: Database deployment layout with KPI increase or decrease relationship

	5 DBs	10 DBs	20 DBs
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [+3%]	997 IOPS [-0%]	948 IOPS [-5%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [+52%]	1686 IOPS [+69%]	1916 IOPS [+92%]
DBs Read Latency [average]	9.9 msec	11.2 msec	17.9 msec
DBs Write Latency [average]	5.6 msec	6.8 msec	8.7 msec
LOGs Write Latency [average]	1.1 msec	1.1 msec	1.6 msec
Resulting latencies after normalization to the Planned Transactional IOPS of 1000 IOPS			
DBs Read Latency normalized	9.6 msec [acting as baseline]	11.3 msec [+18%]	18.8 msec [+97%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	6.8 msec [+25%]	9.2 msec [+69%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	1.1 msec [+5%]	1.7 msec [+60%]

By default, Exchange Server 2010 activates the maintenance tasks on each new mailbox database created. Most of these activities are important to the health and efficiency of the databases and would not be deactivated. Only the online blocks checksum can be optionally offloaded in a scheduled time window if required. The maintenance overhead must be considered when increasing the number of databases, but also carefully considering that large databases offer a lower granularity of administration and data protection.

The results show a uniform trend where the number of databases implemented considerably influence the total IOPS required first, and then the overall function of this kind of deployment. All three criteria report a significant growth up to ten databases, and then further amplify the performance decline in the subsequent test using 20 databases. The database maintenance impact is the measurable element affecting the amount of IOPS dispensed, routinely increasing the percentage of IOPS from 12% to 30%, nearly a third of all the IOPS in the last test case.

4.4 Assess the PS6100 family models

The goal of the PS Series family model assessment analysis was to establish the Exchange KPI trend, IOPS ratios, and their relationship when utilizing different models of EqualLogic PS Series arrays to build the SAN, while maintaining the remaining factors. Table 8 shows the configuration parameters for this test.

Table 8 Test parameters: PS Series model assessment

Reference configuration: Test variables under study	
Number of units, Array model	1x PS6100XV 3.5” or 1x PS6100X or 1x PS6100E
Reference configuration: Consistent factors across this test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Number of simulated mailboxes/users	5,000 concurrent users
Number of databases	5 databases (active)
Mailbox allocation	1,000 mailboxes per each mailbox database
Mailbox size	1GB each
Database size	1TB each
Number of database replica copies	2 (2 node DAG)
RAID policy	RAID 6 / RAID 50
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
SAN configuration	one single pool (default)

The list shows the essential details of the EqualLogic PS6100 used. For additional details refer to the ‘Hardware components’ list in Appendix A.1.

- PS6100XV 3.5”: dual controllers with four GbE ports each, and dedicated management port
  24 3.5” 15,000RPM SAS disk drives, 600GB each
- PS6100X: dual controllers with four GbE ports each, and dedicated management port
  24 2.5” 10,000RPM SAS disk drives, 900GB each
- PS6100E: dual controllers with four GbE ports each, and dedicated management port
  24 3.5” 7,200RPM NL-SAS disk drives, 2TB each

The main difference between the array models within the PS6100 families analyzed during the tests is the drive type. The illustration of the PS6100E in Figure 8 is a stretched analysis to show the benefits and limitations of the platform when compared with the other faster models.

Figure 8 shows the results collected from the Exchange Jetstress simulations when changing the model of the array used to build the EqualLogic SAN.

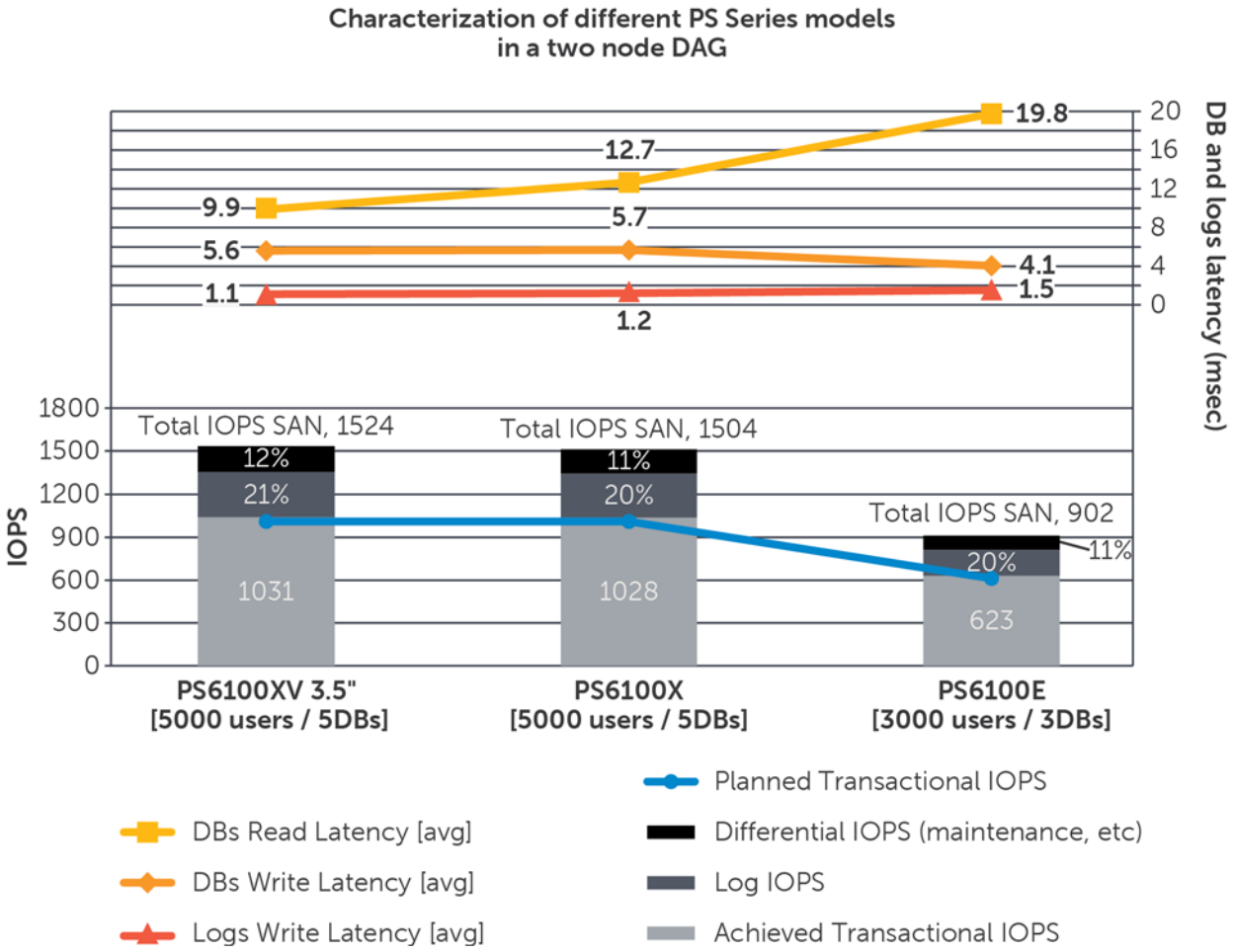

Figure 8 Characterization of different PS Series models in a two node DAG

Table 9 reports the numerical results recorded during the assessment of the PS Series model test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 9 Test results: PS Series model assessment with KPI increase or decrease relationship

	PS6100XV 3.5"	PS6100X	PS6100E
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [+3%]	1028 IOPS [+3%]	623 IOPS [+4%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [+52%]	1504 IOPS [+50%]	902 IOPS [+50%]
DBs Read Latency [average]	9.9 msec	12.7 msec	19.8 msec
DBs Write Latency [average]	5.6 msec	5.7 msec	4.1 msec
LOGs Write Latency [average]	1.1 msec	1.2 msec	1.5 msec
Resulting latencies after normalization to the Planned Transactional of …	1000 IOPS		600 IOPS
DBs Read Latency normalized	9.6 msec [acting as baseline]	12.3 msec [+29%]	19.0 msec [+99%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	5.5 msec [+2%]	3.9 msec [-18%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	1.2 msec [+10%]	1.5 msec [+37%]

The outcomes show the higher rotational speed of the drives is the first aspect that affects the efficiency. The PS6100XV 3.5" equipped with 15K RPM SAS drives leads the performance for the pool of models assessed, followed by the PS6100X with 10K RPM SAS drives, and then the PS6100E with 7.2K NL-SAS drives, where the characteristic random IOPS of Exchange creates a large difference within the read latency indicator. The PS6100E was able to power up to 3,000 users with the same mailbox user profile used on the other array models.

4.5 Scale the SAN and the users

The goal of the scaling up and out analysis was to establish the Exchange KPI trend, IOPS ratios, and their relationship when scaling up the number of concurrent mailbox users and afterward scaling out the SAN, while maintaining the remaining factors. Table 10 shows the configuration parameters for the first set of these tests.

Table 10 Test parameters: Scaling up the number of concurrent mailbox users

Reference configuration: Test variables under study	
Number of simulated mailboxes/users	5,000 / 7,000 / 8,000 concurrent users
Number of databases	5 / 7 / 8 databases (active)
Reference configuration: Consistent factors across this test	
Messages per day per mailbox / IOPS per mailbox	200 messages / 0.20 IOPS (with DAG)
Mailbox allocation	1,000 mailboxes per each mailbox database
Mailbox size	1GB each
Database size	1TB each
Number of database replica copies	2 (2 node DAG)
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
RAID policy	RAID 6
Number of units, Array model, SAN configuration	1x PS6100XV 3.5", one single pool (default)

The building block used to scale up the number of users for the tests was the addition of databases, while maintaining the users per database ratio at 1,000:1. Gradually provisioning predefined-sized mailbox databases simplifies the administrative burden when the demand to support more users increases and conveniently increments the workload in a linear fashion.

The test series includes three different iterations increased by a progressive amount of workload. It scales up the amount of concurrent mailbox users from 5,000, to 7,000, and then to 8,000 by incrementally adding mailbox database units of 1TB containing 1,000 mailboxes each.

The database cache evaluated for the pool of users in the test series is reported in Table 11, again without considering the additional memory requirements due to other factors. The estimates for the database cache are based on Microsoft published metrics and not on recorded values from our tests. Microsoft Exchange Jetstress memory and processor utilization performs differently from a full Exchange Server, as reported in Appendix B.

Table 11 Exchange database cache (estimated from published metrics) while scaling up the number of users

Number of concurrent mailbox users	Cache per user	Exchange database Cache (estimate)
5,000	12MB	60GB
7,000		84GB
8,000		96GB

Figure 9 shows the results collected from the Exchange Jetstress simulations when scaling up the number of mailbox users.

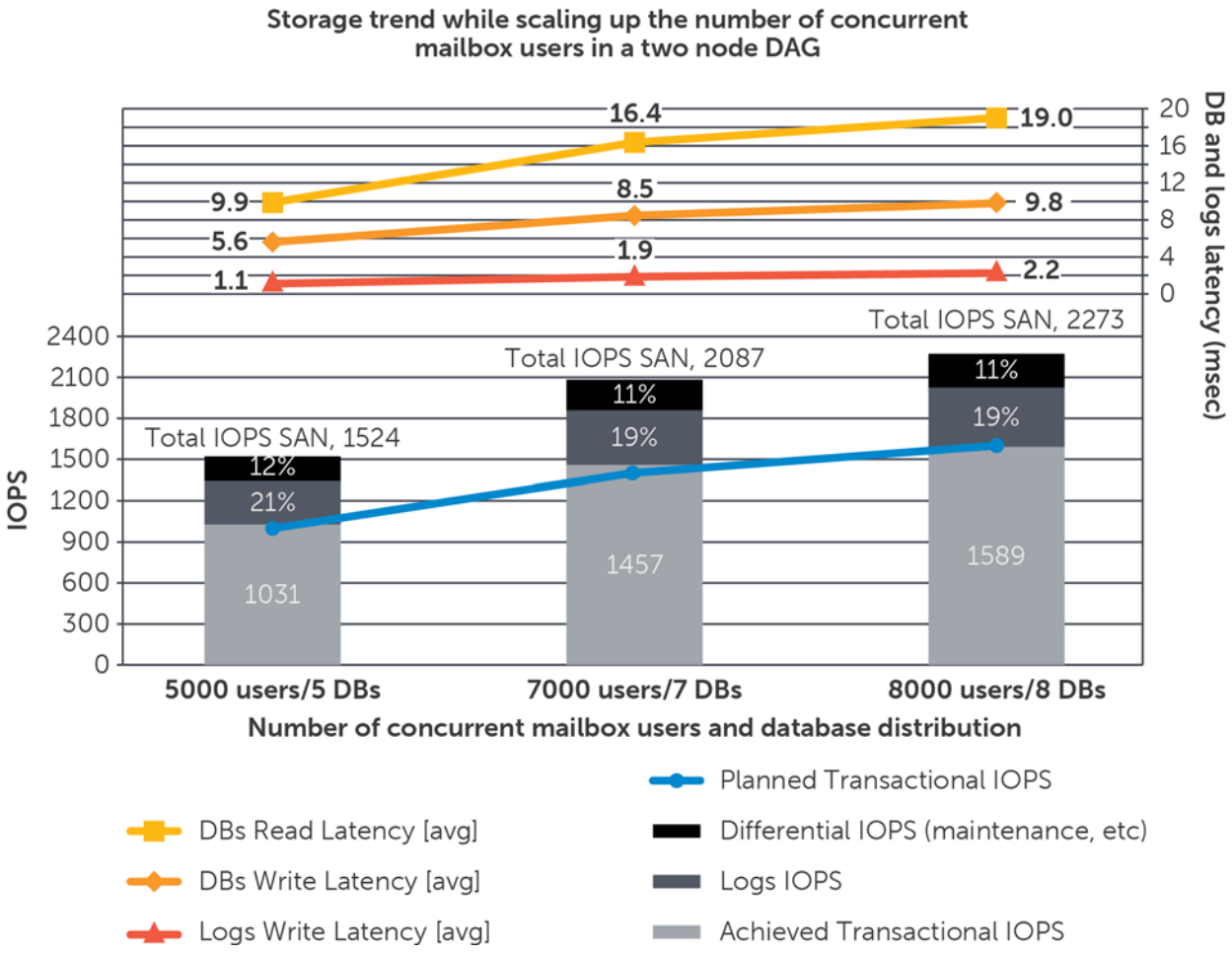

Figure 9 Storage trend while scaling up the number of concurrent mailbox users in a two node DAG

Table 12 reports the numerical results recorded during the increase of the number of concurrent mailbox users test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 12 Test results: Scaling up concurrent mailbox users with KPI increase or decrease relationship

	5,000 users	7,000 users	8,000 users
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [+3%]	1457 IOPS [+4%]	1589 IOPS [-1%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [+52%]	2087 IOPS [+49%]	2273 IOPS [+42%]
DBs Read Latency [average]	9.9 msec	16.4 msec	19.0 msec
DBs Write Latency [average]	5.6 msec	8.5 msec	9.8 msec
LOGs Write Latency [average]	1.1 msec	1.9 msec	2.2 msec
Resulting latencies after normalization to the Planned Transactional of …	1000 IOPS	1400 IOPS	1600 IOPS
DBs Read Latency normalized	9.6 msec [acting as baseline]	15.7 msec [+64%]	19.2 msec [+100%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	5.4 msec [+50%]	9.9 msec [+81%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	1.1 msec [+66%]	2.3 msec [+110%]

The outcomes of this series reveal a nearly linear scalability. In Figure 9 above, you can see the trend of the three latency KPI lines that closely follow the planned IOPS (blue line). Even with the penalty of the increased number of databases, as illustrated in Section 4.3, the latency metrics remain beneath the reference Microsoft thresholds.

The next step taken in the tests was to scale out the building block defined by the reference workload, by one server, and by one SAN array. We ran two series of tests, each executing the three tests illustrated for scaling up the number of concurrent mailbox users, where we multiplied the workload, the hosts, the VMs and the SAN arrays by a factor of two. Table 13 shows the configuration parameters for the two series of tests.

Table 13 Test parameters: Scaling out the SAN by a factor of two

Series	Reference configuration: Test variables under study	
#1 & #2	Number of simulated mailboxes/users	10,000 / 14,000 / 16,000 concurrent users
	Number of databases	10 / 14 / 16 databases (active)
#1	SAN Configuration	one single pool (default)
#2	SAN Configuration	two pools (one array per each pool)
Reference configuration: Consistent factors across this test		
Messages per day per mailbox / IOPS per mailbox		200 messages / 0.20 IOPS (with DAG)
Mailbox allocation		1,000 mailboxes per each mailbox database
Mailbox size		1GB each
Database size		1TB each
Number of database replica copies		2 (2 node DAG)
iSCSI initiator software collocation		Guest initiator (residing in the VMs)
RAID policy		RAID 6
Number of units, Array model		2x PS6100XV 3.5”

The building block to scale out the SAN to support larger environments is framed around the user per database ratio of 1,000:1 defined previously. The four components involved are the EqualLogic SAN array, the host hypervisor, the Exchange mailbox server, and the total workload per server, established from the reference workload of 0.20 IOPS per mailbox. The building block ratio of 1:1:1:1 is conserved across the two series.

Figure 10 shows the results collected from the Exchange Jetstress simulations of the two series of incremental workloads while scaling out the SAN and the server resources.

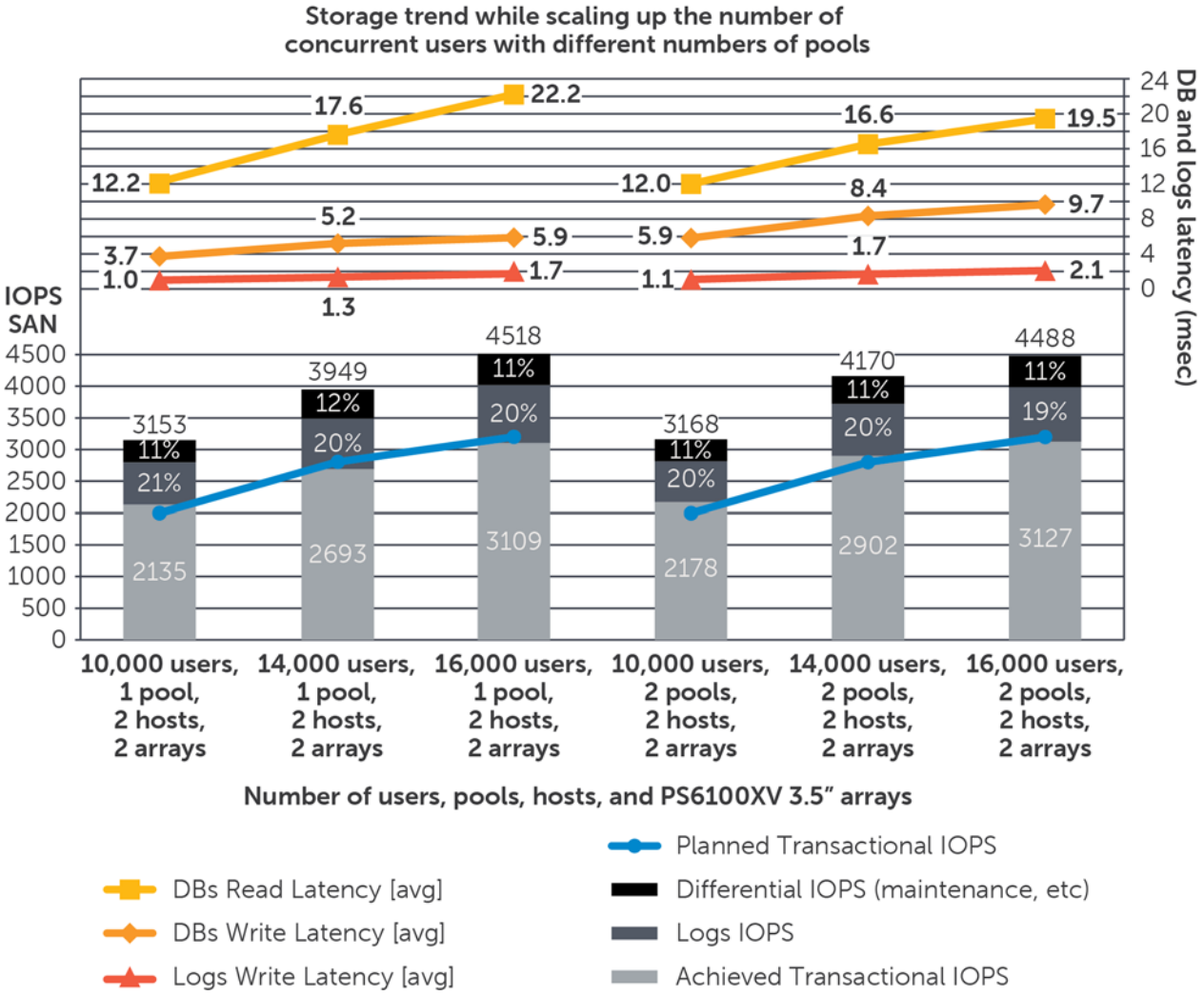

Figure 10 Storage trend while scaling up the number of concurrent users with different number of pools

Table 14 reports the results recorded during the test that increased the number of concurrent mailbox users while scaling out the SAN and the server resources. It also shows the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 14 Test results: Scaling out the SAN by a factor of two with KPI increase or decrease relationship

Series #1	**10,000 users 2 arrays / 1 pool**	**14,000 users 2 arrays / 1 pool**	**16,000 users 2 arrays / 1 pool**
Achieved Transactional IOPS [% different vs. planned IOPS]	2135 IOPS [107%]	2693 IOPS [96%]	3109 IOPS [97%]
Total IOPS of the SAN [% different vs. planned IOPS]	3153 IOPS [158%]	3949 IOPS [141%]	4518 IOPS [141%]
DBs Read Latency [average]	12.2 msec	17.6 msec	22.2 msec
DBs Write Latency [average]	3.7 msec	5.2 msec	5.9 msec
LOGs Write Latency [average]	1.0 msec	1.3 msec	1.7 msec
Resulting latencies after normalization to the Planned Transactional of …	2000 IOPS	2800 IOPS	3200 IOPS
DBs Read Latency normalized	11.4 msec [acting as baseline]	18.3 msec [+60%]	22.9 msec [+100%]
DBs Write Latency normalized	3.5 msec [acting as baseline]	5.4 msec [+55%]	6.0 msec [+73%]
LOGs Write Latency normalized	0.9 msec [acting as baseline]	1.4 msec [+51%]	1.8 msec [+91%]
Series #2	**10,000 users 2 arrays / 2 pools**	**14,000 users 2 arrays / 2 pools**	**16,000 users 2 arrays / 2 pools**
Achieved Transactional IOPS [% different vs. planned IOPS]	2178 IOPS [109%]	2902 IOPS [104%]	3127 IOPS [98%]
Total IOPS of the SAN [% different vs. planned IOPS]	3168 IOPS [158%]	4170 IOPS [149%]	4488 IOPS [140%]
DBs Read Latency [average]	12.0 msec	16.6 msec	19.5 msec
DBs Write Latency [average]	5.9 msec	8.4 msec	9.7 msec
LOGs Write Latency [average]	1.1 msec	1.7 msec	2.1 msec
Resulting latencies after normalization to the Planned Transactional of …	2000 IOPS	2800 IOPS	3200 IOPS
DBs Read Latency normalized	11.1 msec [acting as baseline]	16.0 msec [+45%]	19.9 msec [+80%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	8.1 msec [+50%]	9.9 msec [+84%]
LOGs Write Latency normalized	1.0 msec [acting as baseline]	1.7 msec [+64%]	2.2 msec [+112%]

The outcomes of the both series shows the ability of the EqualLogic SAN to accept and support the additional workload distributed across the two arrays. The distinctive difference of the gathered results underlines the sensitivity of Exchange workloads to read operations.

The single pool configuration, with a shared cache across the arrays, better accommodates the write operations, which usually are the most critical I/O operations in a business environment. When comparing the current results with the ones from the scaling of the number of mailbox users with a single array, at the beginning of section 4.5, we notice a steep decrease for both database and log file write latencies, with the drawback of the rise of database read latency. The two pools layout instead do not dispense any particular gain in any of the areas measured by our KPI, but displays the advantage of preserving the same progression of results as in the tests with a single array, and as such tenders a finer building block for scaling out the Exchange workload on a SAN.

The final test for scaling the Exchange workload was to evaluate the behavior of our selected building block when scaled out up to three modules. We completed two series of tests, the first based on the reference workload of 5,000 concurrent users, the second on the maximum workload assessed of 8,000 concurrent users. Each series comprised one, two and three building blocks of the workload, the hosts, the VMs, and the SAN arrays. Table 15 shows the configuration parameters for the two series of tests.

Table 15 Test parameters: Scaling out from one to three building blocks

Series	**Reference configuration: Test variables under study**	
#1	Number of simulated mailboxes/users	5000 / 10,000 / 15,000 concurrent users
	Number of databases	5 / 10 / 15 databases (active)
	Number of units, Array model, SAN Configuration	1x / 2x / 3x, PS6100XV 3.5”, one / two / three pools
#2	Number of simulated mailboxes/users	8000 / 16,000 / 24,000 concurrent users
	Number of databases	8 / 16 / 24 databases (active)
	Number of units, Array model, SAN Configuration	1x / 2x / 3x, PS6100XV 3.5”, one / two / three pools
Reference configuration: Consistent factors across each test		
Messages per day per mailbox / IOPS per mailbox		200 messages / 0.20 IOPS (with DAG)
Mailbox allocation		1000 mailboxes per each mailbox database
Mailbox size		1GB each
Database size		1TB each
Number of database replica copies		2 (2 node Database Availability Group)
iSCSI initiator software collocation		Guest initiator (residing in the VMs)
RAID policy		RAID 6

Figure 11 shows the results collected from the Exchange Jetstress simulations of the two series of incremental workload while scaling out from one to three building blocks.

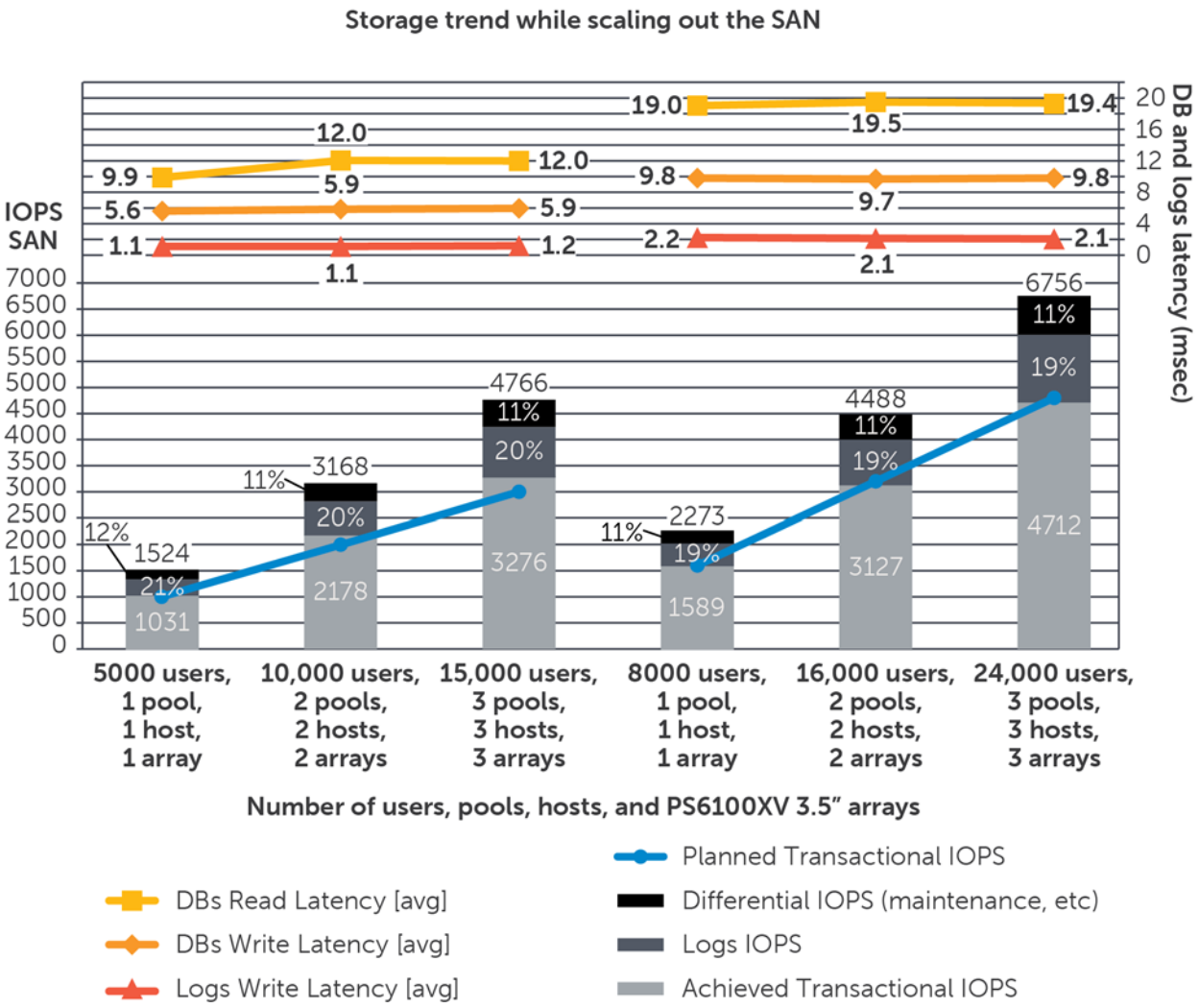

Figure 11 Storage trend while scaling out the SAN

Table 16 reports the numerical results recorded during the increase of building blocks test and the corresponding normalized values to match the planned and achieved IOPS. The percentages reported in the IOPS rows are calculated against the planned IOPS. The rows with the normalized values provide the relative increase or decrease of the Exchange KPI. The responsiveness of the storage subsystem, measured by the database or log latencies, should be regarded as a function of the total amount of operations completed, and not merely the transactional load.

Table 16 Test results: Scaling out from one to three building blocks with KPI increase or decrease relationship

Series #1	**5000 users 1 array / 1 pool**	**10,000 users 2 arrays / 2 pools**	**15,000 users 3 arrays / 3 pools**
Achieved Transactional IOPS [% different vs. planned IOPS]	1031 IOPS [103%]	2178 IOPS [109%]	3276 IOPS [109%]
Total IOPS of the SAN [% different vs. planned IOPS]	1524 IOPS [152%]	3168 IOPS [158%]	4766 IOPS [159%]
DBs Read Latency [average]	9.9 msec	12.0 msec	12.0 msec
DBs Write Latency [average]	5.6 msec	5.9 msec	5.9 msec
LOGs Write Latency [average]	1.1 msec	1.1 msec	1.2 msec
Resulting latencies after normalization to the Planned Transactional of …	1000 IOPS	2000 IOPS	3000 IOPS
DBs Read Latency normalized	9.6 msec [acting as baseline]	11.1 msec [+16%]	11.0 msec [+15%]
DBs Write Latency normalized	5.4 msec [acting as baseline]	5.4 msec [-1%]	5.4 msec [+0%]
LOGs Write Latency normalized	1.1 msec [acting as baseline]	1.0 msec [-5%]	1.1 msec [+2%]
Series #2	**8000 users 1 array / 1 pool**	**16,000 users 2 arrays / 2 pools**	**24,000 users 3 arrays / 3 pools**
Achieved Transactional IOPS [% different vs. planned IOPS]	1589 IOPS [99%]	3127 IOPS [98%]	4712 IOPS [98%]
Total IOPS of the SAN [% different vs. planned IOPS]	2273 IOPS [142%]	4488 IOPS [140%]	6756 IOPS [141%]
DBs Read Latency [average]	19.0 msec	19.5 msec	19.4 msec
DBs Write Latency [average]	9.8 msec	9.7 msec	9.8 msec
LOGs Write Latency [average]	2.2 msec	2.1 msec	2.1 msec
Resulting latencies after normalization to the Planned Transactional of …	1600 IOPS	3200 IOPS	4800 IOPS
DBs Read Latency normalized	19.2 msec [acting as baseline]	19.9 msec [+4%]	19.7 msec [+3%]
DBs Write Latency normalized	9.9 msec [acting as baseline]	9.9 msec [+0%]	10.0 msec [+1%]
LOGs Write Latency normalized	2.3 msec [acting as baseline]	2.2 msec [-4%]	2.1 msec [-6%]

The final outcomes of the building blocks scaling exercise present an outstanding linear scalability of the solution proposed, shown both in the linear chart and in the percentages achieved or normalized always standing around 100%. This shows a limited or null deviation from the behavior of the single building block when we scale it out for a larger environment with heavier workloads.

4.6 Exchange 2010 DAG databases activation and operations

In the second phase of the tests, the assessment of some operative scenarios that will eventually occur in a production environment was undertaken. A fully deployed Exchange 2010 organization must evaluate these elements. For details about the design and configurations, refer to the second part of Section 3.1 and to Appendix A.

The configuration for the mailbox database servers and the activation of the databases are reported in Table 17.

Table 17 Reference configuration for Microsoft Loadgen 2010 tests

Reference configuration	
Number of simulated mailboxes/users	5,000 concurrent users
Number of servers	2 nodes configured in a DAG
Number of databases	10 databases (5 active, 5 passive)
Mailbox allocation	1,000 mailboxes per each mailbox database
Active databases per server	5 DBs / no DB or 4 DBs / 1 DB or 3 DBs / 2 DBs
Active users count per server	5,000u / no users or 4,000u / 1,000u or 3,000u / 2,000u
Exchange Server Cache in use (estimated)	60GB / none or 48GB / 12GB or 36GB / 24GB
Mailbox size	1GB each (with around 34,000 objects per mailbox distributed across folders of less than 5,000 items)
Capacity allocated	Database size (theoretical): 1TB each Database size (recorded): 1.254TB each (average)
	Catalog Index size (recorded): 530GB each database *
	Log files capacity (estimated): 205GB per day (41GB per database), circular logging not enabled
Exchange Search Indexer	Running
Exchange Database Maintenance	Enabled in background (24x7)
RAID policy	RAID 6
iSCSI initiator software collocation	Guest initiator (residing in the VMs)
Number of units, Array model, SAN configuration	2x PS6100XV 3.5”, two pools, one pool per each server
Windows Disk/Partition, File System	Basic disk, GPT partition, default alignment NTFS, 64KB allocation unit size

*The usual average size of an Exchange 2010 Catalog Index is 5-10% of the size of the database. A side effect of the population of Exchange data by Loadgen tool is the exponential growth of the indexes, 40-45% in our case (see Appendix B for details).

Exchange Server cache (estimated) minimum requirement per each DAG node is calculated by

$$Memory = minimum\ RAM\ with\ 1to10\ DBs + (\#users * Cache\ per\ user) = 2GB + (5{,}000 * 12MB) = 62GB$$

and thus requires a building block server or VM with at least 64GB of RAM.

Mailbox databases size is calculated by the combination of the following formulas

$$Dumpster\ delta = Dumpster\ size + Single\ item\ recovery + Calendar\ logging = 210 + 12 + 30 = 252MB$$

$$Dumpster\ size = \#Sent\ \&\ received\ msgs\ per\ user\ per\ day * Avg\ msg\ size * Deleted\ item\ retention\ time = 200 * 75KB * 14 = 210MB$$

$$Single\ item\ recovery = Mailbox\ storage\ quota * 0.012 = 1GB * 0.012 = 12MB$$

$$Calendar\ logging = Mailbox\ storage\ quota * 0.03 = 1GB * 0.03 = 30MB$$

$$White\ space = \#Sent\ \&\ received\ msgs\ per\ user\ per\ day * Avg\ msg\ size = 200 * 75KB = 15MB$$

and the database size expected is very close to the capacity allocated values recorded in Table 17 above.

$$Mailbox\ DB\ size = Number\ of\ mailboxes * (Mailbox\ storage\ quota + Dumpster\ Delta + White\ Space) = 1{,}000 * (1GB + 252MB + 15MB) = 1.267TB$$

Transaction log files (estimated) requirement is calculated by

$$\#\ log\ files = \#users * \frac{2.73 * \ \#Sent\ \&\ received\ msgs\ per\ user\ per\ day * Avg\ msg\ size * Msg\ size\ factor}{Size\ of\ log\ \ (1MB)}$$

$$= 1{,}000 * \frac{2.73 * \ 200 * 75KB * 1}{1MB} = 1{,}000 * 41 = 41{,}000\ or\ 41GB\ per\ day$$

Note: For additional information about the above formulas and sizing exercises, refer to 'Capacity planning and sizing' section of the whitepaper *Sizing Microsoft Exchange 2010 on EqualLogic PS6100 and PS4100 Series Arrays with VMware vSphere 5* http://en.community.dell.com/techcenter/storage/w/wiki/3626.sizing-microsoft-exchange-2010-on-equallogic-ps6100-and-ps4100-series-arrays-with-vmware-vsphere-5-by-sis.aspx

Databases activation (options) is any combination from zero to five active databases on either of the two node DAG servers (the mailbox number allocation and estimate Exchange cache is reported in Table 17 above)

1. Server A owns five active databases, Server B owns the five passive copies and is considered in full standby
2. Server A owns four active and one passive databases, Server B owns one active and four passive (or vice versa)
3. Server A owns three active and two passive databases, Server B owns two active and three passive (or vice versa)

A list of failure events were identified that can affect the HA level of the messaging operations in this infrastructure and the options to address those scenarios were investigated.

- In the event of a failure of either of one of the servers, all the active databases will be switched over to the remaining working server (five active), without any replicated passive copies available.
- In the event of a failure of either one of the SAN storage pools, again the five databases will be activated on the server with access to the functioning storage pool.
- In the event of a database corruption, or an unforeseen administrative fault that terminates with the loss of one or more database copies, the healthy copies remaining will be activated.

The solution steps to reestablish the expected level of availability (considered to be when two consistent copies of each database are available, one active and one passive) are reassumed in the following conceptual operations list:

1. Rebuild or restore the Exchange server lost (in case of loss of server)
2. Rebuild or replace the SAN storage pool (in case of loss of one pool)
3. Reinitialize and connect the (new) storage volumes to the (new) server
4. Join or verify the membership of the (new) server to the DAG
5. Proceed with a DAG seeding operation of the number of lost databases: this could be from one copy only, in the case of a single database corruption, to up to five copies in the case of server or storage pool loss

Figure 12 reports the elapsed time to replicate one to five databases, as the ones listed in Table 17 above, including their own Catalog Index with one or more seeding operations running in parallel, with the aim to reestablish the original level of data availability between the nodes in the DAG.

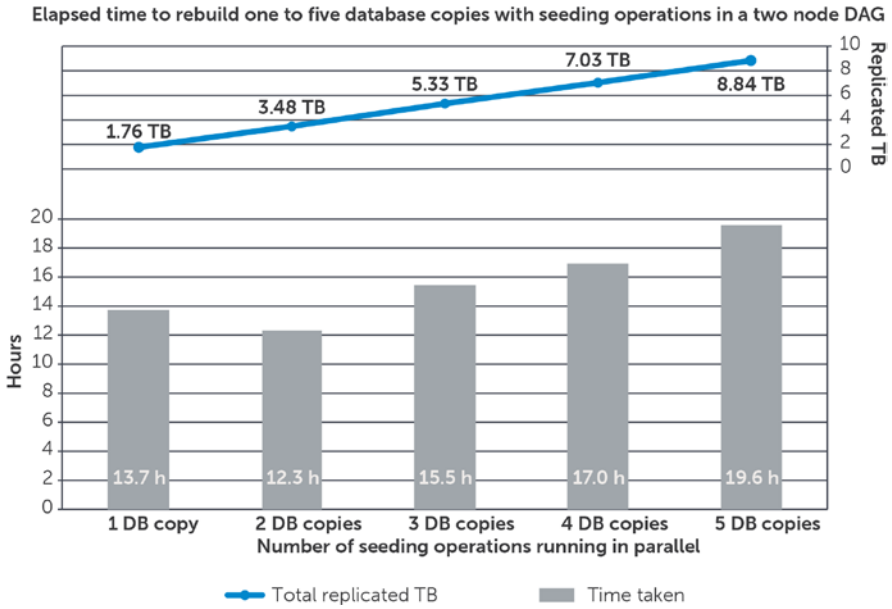

Figure 12 Elapsed time to rebuild one to five database copies with seeding operations in a two node DAG

The time taken to complete these tasks corresponds to the amount of data to be replicated, though it is not duplicate or triplicate like the growth of the number of seeding operations.

The first question to evaluate is if the duration of any of these sets of seeding operations is satisfactory for the service level agreement of the messaging system in the organization. The loss of any of the databases in an Exchange DAG with two copies of each of them represents a potential threat to the production environment until the second copy is reinstated. To bring back the DAG to its data redundancy status it is vital to assess and minimize the seeding time of one or more failed databases.

The seeding operations additionally have a resource impact on both sending and receiving servers while the data is transferred. The Exchange replication service is the service responsible for managing the replication at both ends (*msexchangerepl.exe* process). It instructs the Exchange Store service on the node with the active copy to read from the source database, which might likely be under load in a 24x7 working environment, and then on the node hosting the target copy to write. Both the processor utilization and the disk activities have a measurable footprint that climbs with the increment of the number of parallel copies.

Note: For a detailed description and analysis of seeding behavior and the tactics to confront it refer to 'Restore the high availability level in a DAG' section of the whitepaper *Best Practices for Enhancing Microsoft Exchange Server 2010 Data Protection and Availability using Dell EqualLogic Snapshots* *http://en.community.dell.com/techcenter/storage/w/wiki/2633.enhancing-microsoft-exchange-server-2010-data-protection-and-availability-with-equallogic-snapshots-by-sis.aspx*

Therefore to relieve the DAG node server with the active databases from the burden of the replication activities, while simultaneously supporting all the remaining active users, two different approaches can be pursued.

- Add a third node to the DAG and preemptively redistribute the databases across all the nodes. This method allows you to seed one or more lost database copies from another passive copy, while the active copy is left relieved of the Exchange replication flow.
- Protect the mailbox database volumes with EqualLogic Smart Copy technologies and proceed with fast recovery and reseeding directly supported by the SAN. All the steps and options to carry out this solution are explained in the whitepaper referred to in the note box above.

5 Best practices recommendations

Refer to these best practices to plan and configure Hyper-V servers, Exchange Server 2010, and EqualLogic arrays.

Storage best practices

- Use Multipath I/O (MPIO) Device Specific Module (DSM) provided by EqualLogic HIT Kit to improve performance and reliability for the iSCSI connections.
- Distribute network port connections on the controllers accordingly with the port failover mechanism and the redundancy implemented on the network switches.
- Maintain a 1:1 ratio between the number of network ports on the active arrays controller and the number of host network adapters to maximize the utilization of the available bandwidth.
- Carefully choose the most appropriate RAID policy from the beginning according to the performance, capacity, and tolerance to failure requirements of your environment.
- RAID conversion provides outstanding benefits when strictly required. Do not overuse it if a clean configuration is viable. RAID conversions of large volumes take a considerable amount of time and leave the resulting set of data pages extremely fragmented with a potential risk of lower efficiency.
- Do not share the disk drives for active and replicated copies of an Exchange mailbox database. The failure of a set of drives with multiple database copies decreases the resilience or the perceived availability in this deployment scenario. Dedicate separate pools for databases connected to different nodes in a DAG instead.

Network best practices

- Design separated network infrastructures to isolate the LAN traffic from the SAN traffic (iSCSI)
- Implement redundant components (switches, ISLs, network adapters) to provision a resilient network infrastructure between the endpoints (stack, LAG, load balancing or network card teaming)
- Disable spanning tree for the switch ports hosting the PS Series array controllers connections and enable 'Portfast' instead
- Enable flow control for the switch ports hosting the PS Series array controller connections
- Enable flow control on the host network adapters dedicated to SAN traffic (iSCSI)
- Enable jumbo frames (large MTU) for the switch ports hosting the PS Series array controller connections
- Enable jumbo frames on the host network adapters assigned to SAN traffic (iSCSI)
- Evaluate jumbo frames (large MTU) for the LAN network when appropriate (limited by the type of devices the traffic traverses)
- Enable Large Send Offload, TCP, and UDP Checksum Offload for both Rx and Tx on the host network adapters connected to the SAN traffic (iSCSI)

Hyper-V and VMs best practices

- The option of installing a Windows Server Core version in the root partition of the Hyper-V role server is advised when reducing the maintenance, the software attack surface, the memory, and disk space footprint are critical requirements. 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.
- Exchange Mailbox role servers are characterized by a memory intensive workload. Configure static memory in the settings of each VM to avoid allowing the dynamic memory management to create contention between different VMs running on the same host, which could possibly penalize the Exchange storage I/O execution.
- The use of Non-Uniform Memory Access is advised to address the management of VMs with large and very large memory settings. Plan for the configuration of affinity between the NUMA nodes and the VMs depending on the number of either one, available via Windows Management Instrumentation (WMI) scripting.
- Carefully plan the capacity of the volumes hosting the VMs VHD files, including the space required for memory file content (.bin), save state, or snapshot files (.vsv). Remember that differencing VHD files or snapshots of guest VMs hosting Exchange Server 2010 are currently not supported by Microsoft.
- Isolate the host management traffic from the VM traffic preferring virtual switches not enabled for management.
- Traffic segregation for different kind of traffic from your VMs requires corresponding virtual switches, thus host network adapters, to isolate the traffic.
- Avoid mixing LAN and iSCSI traffic on the same virtual adapters: enforced as a consequence of the LAN and iSCSI network isolation design.
- Select VM network bus adapters with synthetic drivers as opposed to legacy network adapters with emulated drivers.
- Configure a dedicated virtual switch for each guest network adapter connected to the SAN traffic (iSCSI) you plan to have in the VM. Maintain a 1:1 ratio between the number of network ports on the active arrays controller and the number of host/guest network adapters configured on the VMs.
- Aggregate at least two host network adapters in failover teams for each virtual switch in order to achieve resiliency
- Enable jumbo frames on both the guest network adapters and the corresponding host adapters assigned to SAN traffic (iSCSI) and to the replication traffic of the Exchange Replication service.
- Enable Large Send Offload, TCP and UDP Checksum Offload for both Rx and Tx on the guest network adapters connected to the SAN traffic (iSCSI).
- Evaluate jumbo frames for the guest network adapters and the corresponding host adapters assigned to LAN traffic.

Exchange installation best practices

- Use Windows 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 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.
- Deploy Windows operating system and Exchange data in physically separated disk drives.
- 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 your DAG configuration.

Know your workload

As simple as it is, do not begin a deployment without having a solid understanding of your current messaging workload. In the case of a greenfield deployment, collect estimates based on business cases matching your organization size and drive conservative figures for the average user profiles.

Project the workload differential between the current and the target releases of Exchange Server when you plan to design a storage solution jointly with a migration.

Exchange mailbox database layout

The right balance of the number of mailbox databases to support a number of users is mostly based on administrative policies. Bigger databases fitted with a large amount of users perform better than an extensive number of small databases because of the aggregation of overhead loads such as database maintenance. Microsoft Windows operating system supports a precise maximum amount of iSCSI targets and connections: do not pass the limits to avoid an unsupported deployment scenario.

Number of mailboxes per mailbox database

Reducing the number of user mailboxes per database provides a more agile environment to administer when using a traditional backup application or even when administrative tasks require you to temporarily dismount the database causing a downtime. The balance between number of databases and user mailboxes per database must be cautiously planned according with the needs and size of the entire Exchange organization. Usually a single database containing a high percentage of the entire organization is perceived as a single point of failure or at least as an administrative constraint.

Understand and clarify RTO and RPO

Define and clarify the RTO and RPO values before designing a highly available solution. The number of cluster nodes or the amount of storage and features available on the SAN greatly affect the flexibility of the topology and deployment selected.

For general recommendations and information about EqualLogic PS Series array configurations, refer to Dell EqualLogic Configuration Guide, available at:
http://www.delltechcenter.com/page/EqualLogic+Configuration+Guide

A Configuration details

A.1 Hardware components

Table 18 lists the details of the hardware components used for the configuration setup.

Table 18 Hardware components

Test configuration – Hardware components	
Servers	Dell PowerEdge M1000e Blade enclosure • 2x Chassis Management Controller (CMC), Firmware 4.11.A01 Dell PowerEdge M710 Blade Server • 2x Quad Core Intel® Xeon X5570 Processors, 2.93 Ghz, 8M Cache • RAM 96 GB (12x 8GB) • PERC 6/I RAID controller • 4x 146 GB 15K SAS (2x RAID-1) • 4x Broadcom NetXtreme II 5709S Dual Port 1GbE Mezzanine Card, Firmware 7.2.2 Dell PowerEdge M610 Blade Server • 2x Quad Core Intel Xeon E5520 Processors, 2.26 Ghz, 8M Cache • RAM 32 GB (8x 4GB) • PERC 6/I RAID controller • 2x 146 GB 15K SAS (RAID-1) • 2x Broadcom NetXtreme II 5709S Dual Port 1GbE Mezzanine Card, Firmware 6.4.5
Network	2x Dell PowerConnect M6220 Blade Switches (stacked), Firmware 4.2.2.3 • installed in M1000e blade enclosure fabrics A1 and A2 2x Dell PowerConnect M6348 Blade Switches (vertically stacked in pairs with the 7048R) , Firmware 4.2.2.3 • installed in M1000e blade enclosure fabrics B1 and B2 2x Dell PowerConnect 7048R Switches (vertically stacked in pairs with the M6348) , Firmware 4.2.2.3 • installed top of the rack
Storage	3x Dell EqualLogic PS6100XV 3.5” • Dual 4-port 1GbE controllers, Firmware 5.2.5 • 24x 600GB 15K 3.5” SAS disk drives, raw capacity 14.4 TB each 1x Dell EqualLogic PS6100X • Dual 4-port 1GbE controllers, Firmware 5.2.5 • 24x 900GB 10K 2.5” SAS disk drives, raw capacity 21.6 TB 1x Dell EqualLogic PS6100E • Dual 4-port 1GbE controllers, Firmware 5.2.5 • 24x 2TB 7.2K 3.5” NL-SAS disk drives, raw capacity 48 TB

A.2 Software components

The setup of the environment required to run the tests reported in this paper included the following software components:

- Hypervisor: Windows Server 2008 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 2008 R2 on every VM
- Dell EqualLogic Host integration Toolkit to provide Dell MPIO access to the back-end SAN on each virtual or physical machine access
- Dell EqualLogic SAN Headquarters to monitor the health and performance of the SAN
- Microsoft Exchange Jetstress to simulate the access to the storage subsystem for all MBX pre-fixed VMs
- Microsoft Exchange Server 2010 to run the fully populated messaging environment for all management, CAS, HUB and mailbox server role installed in the organization
- Microsoft Exchange Loadgen to simulate the client side access to the Exchange infrastructure for the LOAD VM

The following software components were installed and configured to support the Microsoft Exchange organization and to simplify the management of the environment:

- Active Directory Domain Services and DNS Server roles for the domain controller VM
- Microsoft System Center 2012 Virtual Machine Manager (SCVMM) for the management VM (not strictly required to accomplish the tests)

Table 19 lists the details of the software components used for the configuration setup.

Table 19 Software components

Test configuration – Software components	
Operating Systems	Host: Microsoft Windows Server 2008 R2 Enterprise Edition Service Pack 1 (build 7601) with Hyper-V • Dell OpenManage Server Administrator 7.1.0 • Broadcom Advanced control Suite 4 (version 15.2.20.2) • Dell EqualLogic Host Integration Toolkit 4.0 (for host initiator tests only) Guest: Microsoft Windows Server R2 Enterprise Edition Service Pack 1 (build 7601) • Hyper-V Integration Services • Dell EqualLogic Host Integration Toolkit 4.0 (for all guest initiator tests) • MPIO enabled using EqualLogic DSM for Windows
Applications	Microsoft Exchange Server 2010 Enterprise Edition Service Pack 2 (version 14.2, build 247.5) Microsoft Filter Pack 2.0 Microsoft System Center 2012 Virtual Machine Manager (version 3.0.6005.0)
Monitoring tools	Dell EqualLogic SAN Headquarters 2.2 (build 2.2.0.5924) Microsoft Performance Monitor from the Windows Operating System Performance and Mail Flow tools from Microsoft Exchange Management Console
Simulation tools	Microsoft Exchange Jetstress 2010 (build 14.01.0225.017) • Exchange 2010 Server Database Storage Engine and Library Service Pack 2 (build 14.01.0218.012) Microsoft Exchange Loadgen 2010 (build 14.01.0180.003)

A.3 Network configuration details

Two physical networks were built in order to provide full isolation between regular IP traffic and iSCSI data storage traffic. Also, each IP network was segregated from the others by the use of VLANs with tagged traffic. In order to achieve network resiliency for hardware faults, at least two physical switches were stacked for each network, as well as used redundant uplinks (LAG) between the stacked switches. Some important configuration aspects were:

- Flow control enabled for every switch port on 7048Rs and M6348s

- Spanning tree Portfast enabled for every switch port on 7048Rs and M6348s
- Jumbo frames enabled for every switch port on 7048Rs, M6348s and M6220

Table 20, Table 21, and Table 22 summarize the different aspects of the networks implemented in the reference configuration and their usage.

Table 20 Configuration - Switch modules

Switch Module	Placement	Purpose
PowerConnect M6220 #1	M1000e I/O Module A1	Regular IP traffic
PowerConnect M6220 #2	M1000e I/O Module A2	Regular IP traffic
PowerConnect M6348 #1	M1000e I/O Module B1	iSCSI data storage traffic
PowerConnect M6348 #2	M1000e I/O Module B2	iSCSI data storage traffic
PowerConnect 7048R #1	Top of the rack	iSCSI data storage traffic
PowerConnect 7048R #2	Top of the rack	iSCSI data storage traffic

Table 21 Configuration – Host to switch connections

Server	Interface	Number of NIC ports	Purpose
PowerEdge M710	LOM on A1	2	Regular IP traffic
	LOM on A2	2	
	Mezzanine Card on B1	2	iSCSI data storage traffic
	Mezzanine Card on B2	2	
	Total ports = 8 (4 on Fabric A, 4 on Fabric B)		
PowerEdge M610	LOM on A1	1	Regular IP traffic
	LOM on A2	1	
	Mezzanine Card on B1	1	iSCSI data storage traffic
	Mezzanine Card on B2	1	
	Total ports = 4 (2 on Fabric A, 2 on Fabric B)		

Table 22 Configuration - VLANs

VLAN ID	Switch it is implemented on	Purpose
100	PowerConnect M6220	Management
200	PowerConnect M6220	LAN traffic
1 (default)	PowerConnect M6348, PowerConnect 7048R	iSCSI

Figure 13 illustrates the diagram of the connectivity between the two pairs of network switches and between them and the storage arrays. Each top of the rack PC7048R switch is interconnected with one IOM M6348 by two vertical uplink stack, additionally one ISL LAG connects in a mesh pattern the expansion modules from each of the top of rack switches and IOM switches.

For a more detailed and comprehensive approach to the connectivity between IOMs and top of rack switches refer to the whitepaper *SAN Design Best Practices for the Dell PowerEdge M1000e Blade Enclosure and EqualLogic PS Series Storage (1GbE)*, available at: http://en.community.dell.com/techcenter/storage/w/wiki/3893.san-design-best-practices-for-the-dell-poweredge-m1000e-blade-enclosure-and-equallogic-ps-series-storage-1gbe-by-sis.aspx

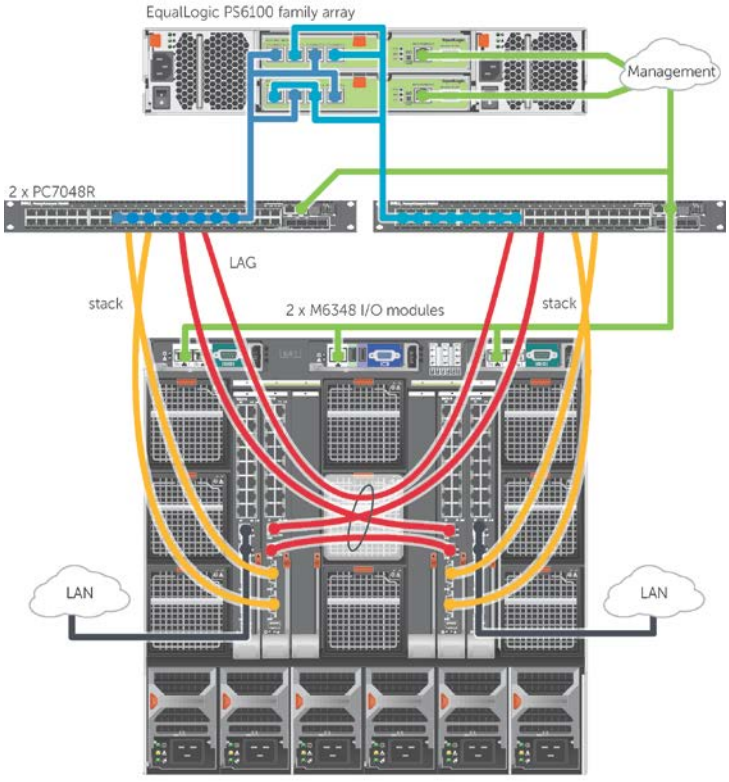

Figure 13 Top of the rack and IOM switches stack diagram

Figure 14 presents the diagram of the network connections between the blade servers and the storage arrays.

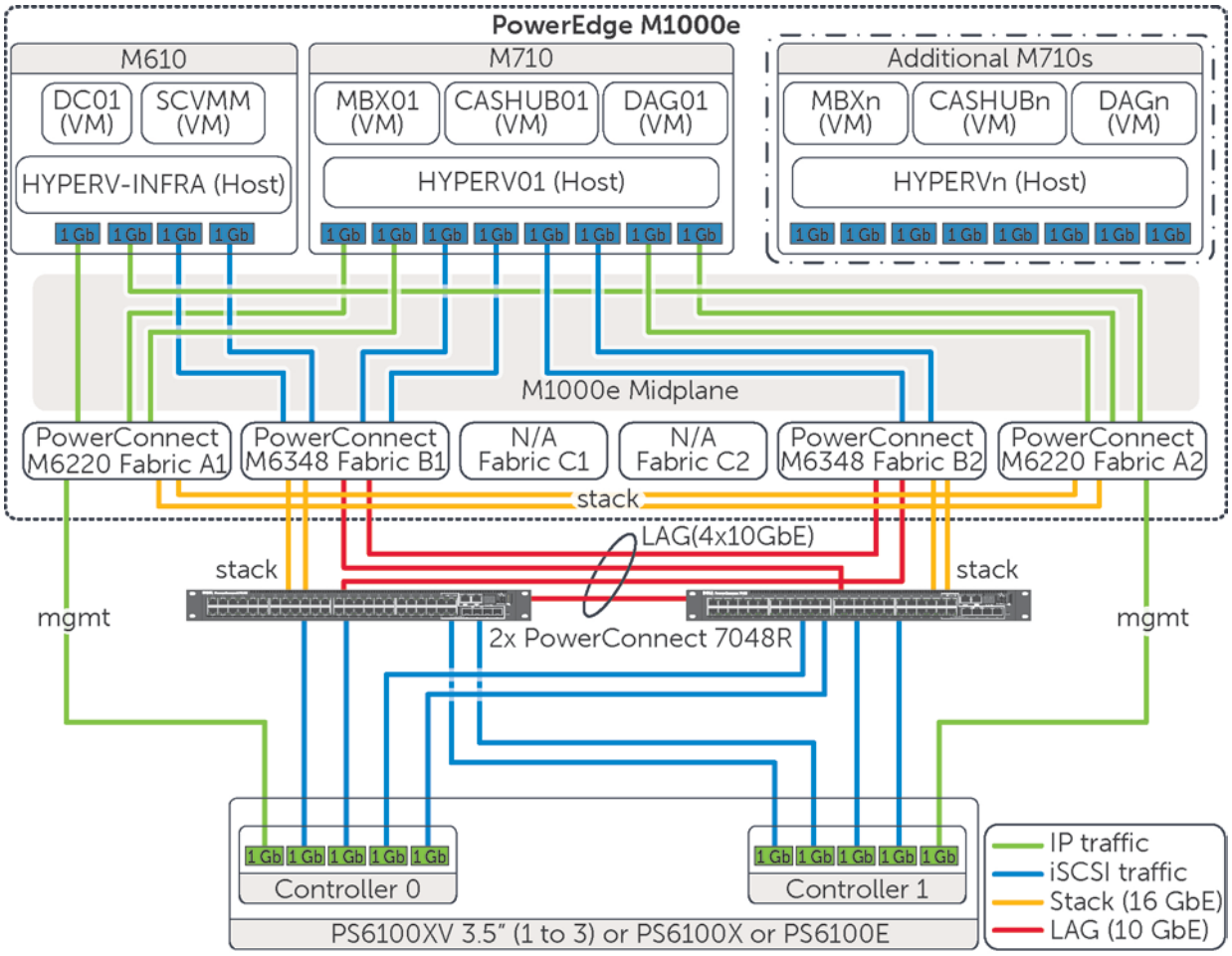

Figure 14 Network connectivity diagram

A.4 Host hypervisor and virtual machines configuration

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

- Windows Server 2008 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 (M710s) configured as member server of the domain
- Two hypervisor hosts (M610s) configured as member server of the domain
- All guests deployed from a single image template of the Windows Server 2008 R2 operating system
- Guest VM operating system disks statically deployed into each host local disks/partitions

- Guest (or host, when described) iSCSI initiator used to access volumes hosted on the EqualLogic SAN

Table 23 lists the relation between the hypervisor host and each VM, with a brief summary of the virtual resources allocated for each VM.

Table 23 Configuration – Host and guest allocation

Host	VM	Purpose	vCPUs	Memory	Network Adapters
HYPERVINFRA (M610)	DC01	Active Directory Domain Controller	2	4GB	1x VM Bus Network Adapter*
	SCVMM	System Center Virtual Machine Manager	2	8GB	2x VM Bus Network Adapters*
HYPERVLOAD (M610)	LOAD	Loadgen simulation	4	16GB	1x VM Bus Network Adapter*
HYPERVn (M710s)	MBXn	Exchange Server Mailbox role simulation (Jetstress)	2	8GB	5x VM Bus Network Adapters*
	CASHUBn	Client Access Server role HUB Server role	4	8GB	2x VM Bus Network Adapters*
	DAGn	Mailbox Server role and DAG node	4	64GB	6x VM Bus Network Adapters*

* All the network adapters implemented in the virtual machines used VM Bus network adapters with synthetic drivers as opposed to legacy network adapters with emulated drivers.

Guest VMs startup allocation

Although the memory and processor size of the servers running the hypervisor were able to run all the VMs at once, we defined the following schedule for the resource allocation of the VMs:

- Jetsress test cases: MBXn VMs started and in use, DAGn/CASHUBn VMs turned off
- Loadgen test cases: DAGn/CASHUBn VMs started and in use, MBXn VMs turned off

Guest VMs memory

The memory assigned to every VM in the infrastructure was 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 physical host (Node Interleaving disabled in the server BIOS) to allow memory access across CPUs. The Hyper-V NUMA Spanning setting was enabled as shown in Figure 15.

The PowerEdge M710s used for our tests had two NUMA nodes each managing 48GB of memory. In order to run a single VM with an amount of memory greater than 48GB the NUMA activation was required (i.e. Exchange DAG nodes have 64GB each).

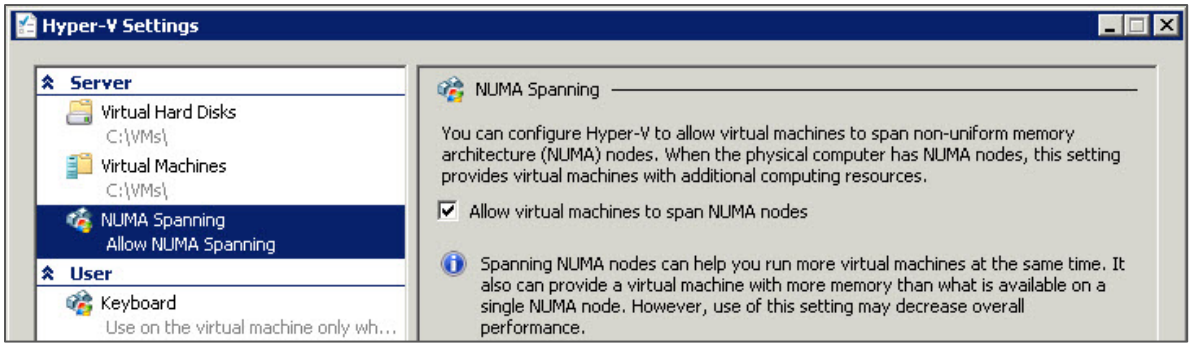

Figure 15 Hyper-V setting for NUMA spanning

Guest VMs disks

The virtual disks in use to host the operating system of each VM were VHD fixed type disks, similar to the configuration shown in Figure 16 and were deployed on the second pair of RAID-1 disks on each host hypervisor.

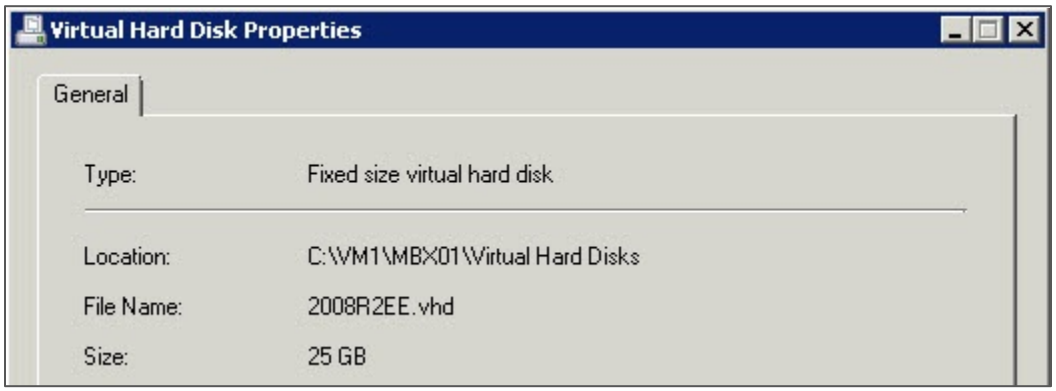

Figure 16 Fixed size VHD files

During the host initiator test case the SAN volumes were presented to the hypervisor host as opposed to directly to the guest VMs. Figure 17 shows how the VHD disks were distributed across them.

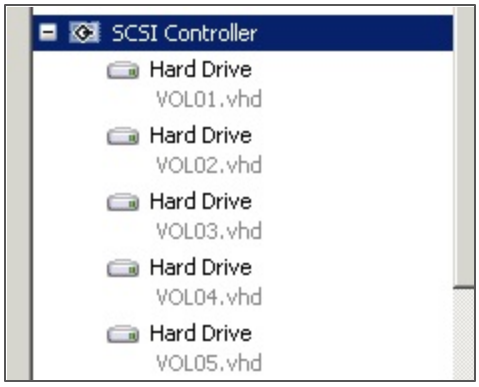

Figure 17 Enumeration of VHD disks for the host initiator test case

A.4.1 Host network adapters and Virtual networks configuration

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

- Two physical network adapters, respectively sourced from Fabric A1 and A2, aggregated in one Broadcom TM/SLB network team (Smart Load Balancing and Failover), and providing connectivity for host domain access and management
- Two physical network adapters, respectively sourced from Fabric A1 and A2, aggregated in one Broadcom TM/SLB network team (Smart Load Balancing and Failover), and providing connectivity for VMs, both domain and intra-VM traffic
- Two (on the M610s) or four (on the M710s) physical network adapters, sourced from Fabric B1 and B2, individually connected to the iSCSI network, and providing access to the SAN from the host or guest environments depending on the test case configuration (host or guest initiator)

Figure 18 shows the detailed hierarchy of the network adapters aggregated in teams for one of the M710s.

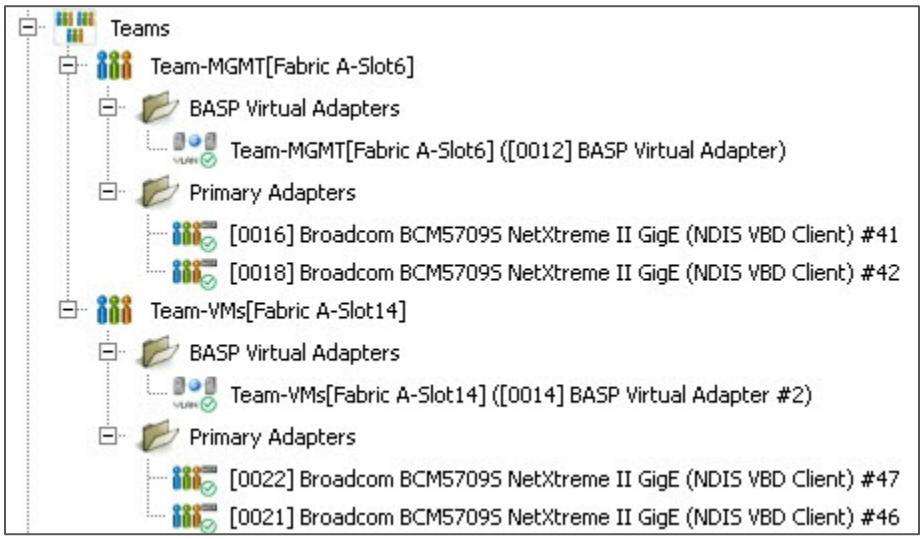

Figure 18 Network teams in use on one of the M710s

The virtual networks configured for the Hyper-V hypervisor servers followed the guidelines listed below.

- One virtual switch with external connection type and no management defined to provide access from the VMs to the LAN traffic – all hosts
- One virtual switch with external connection type and no management defined to provide access from the VMs to the management network – only for the HYPERVINFRA host
- Four virtual switches with external connection type and no management defined to provide access from the VMs to the iSCSI network for the test cases of guests initiator – all HYPERVn hosts

A.4.2 Virtual network adapters configuration

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

- One virtual network adapter to access the Corporate or 'LAN traffic' VLAN, for each VM
- One virtual network adapter to access the dedicated Exchange replication subnet, for the Exchange mailbox servers participating in a DAG, with Jumbo Frames enabled
- One virtual network adapter to access the dedicated management subnet, for the SCVMM VM

Additional network adapters connected to the iSCSI network to access the SAN during the guest initiator test cases, and relevant configurations.

- Four virtual network adapters to access the iSCSI network, for the Exchange mailbox servers or for the Jetstress simulators
- MPIO (Multi-path I/O) enabled and provided by EqualLogic DSM module
- Jumbo frames enabled
- Large Send Offload enabled, TCP and UDP Checksum Offload enabled for both Rx and Tx

Table 24 summarizes the assignment of virtual network adapter to each network grouped by VM.

Table 24 Configuration – VMs network adapters

VM	#NetAdapter	Adapter Type	Virtual Networks	Physical adapters	VLAN ID
DC01	#1	VM Bus	vNet1	Individual	200
SCVMM	#1	VM Bus	vNet1	Individual	200
	#2	VM Bus	vNet0	Individual	100
MBXn	#1	VM Bus	vNet1	Team VMs	200
	#2	VM Bus	vNet-iSCSI0	Individual	NONE
	#3	VM Bus	vNet-iSCSI1	Individual	
	#4	VM Bus	vNet-iSCSI2	Individual	
	#5	VM Bus	vNet-iSCSI3	Individual	
CASHUBn	#1	VM Bus	vNet1	Team VMs	200
	#2	VM Bus	vNet1	Team VMs	200
DAGn	#1	VM Bus	vNet1	Team VMs	200
	#2	VM Bus	vNet1	Team VMs	200
	#3	VM Bus	vNet-iSCSI0	Individual	NONE
	#4	VM Bus	vNet-iSCSI1	Individual	
	#5	VM Bus	vNet-iSCSI2	Individual	
	#6	VM Bus	vNet-iSCSI3	Individual	

B Simulation tools

B.1 Microsoft Jetstress considerations

Microsoft Exchange Server Jetstress 2010 is a simulation tool able to reproduce the database and logs I/O workload of an Exchange mailbox database 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 full Exchange Server 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 (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 per database and SluggishSessions threads property (fine tuning for threads execution pace)

B.2 Microsoft Load Generator considerations

Microsoft Exchange Load Generator 2010 is a validation and stress tool able to generate client access workload against an Exchange 2010/2007 infrastructure. The tool simulates the access patterns of common client applications or protocols such as Microsoft Office Outlook 2007/2003 (online or cached), POP, IMAP4, SMTP, ActiveSync, and Outlook Web App. Some elements worth being considered about Microsoft Exchange Loadgen are:

- Requires a fully functional deployment of a Microsoft Exchange Server 2010 organization
- Requires and performs an initialization step to create Active Directory user accounts and mailbox objects, before it populates the mailboxes according to the defined requirements and settings
- Simulates user client tasks within a wide range: logon/logoff, browse calendar/contacts, create tasks/appointments, send emails, read and process messages, download Outlook Address Book (OAB), etc.
- Reports a high level pass/fail metric for each simulated run (based on the number of the tasks planned to be executed, tasks achieved, and length of the tasks queue maintained)
- Exchange 2010 Content Index size is usually 5-10% of its own mailbox database. It is a known behavior that the size of catalogs created upon mailbox databases built by Loadgen can grow over the expected percentage statistic reported above. The root cause for that has to be seek within the message mix generation and consequent mailboxes population executed by the tool:
  - Loadgen populates the mailbox databases from a predefined small set of emails
  - Email content is an artificial Latin mix with few words concatenated
  - Little variety in the attachments, which are mostly text based and searchable, is used. Some real world attachments are instead not indexable, thus not searchable.

For additional information about Microsoft Exchange Jetstress and Exchange Loadgen 2010 refer to Microsoft documentation: *Tools for performance and Scalability Evaluation*, available at: http://technet.microsoft.com/en-us/library/dd335108.aspx

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 your installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Configuration Guide http://www.delltechcenter.com/page/EqualLogic+Configuration+Guide
- Dell EqualLogic PS Series Group Administration Guide https://support.equallogic.com/support/download_file.aspx?id=1125
- Dell PowerEdge Blade Server and Enclosure Documentation http://support.dell.com/support/edocs/systems/pem/en/index.htm
- Dell PowerConnect 70xx Documentation http://support.dell.com/support/edocs/network/PC70xx/en/index.htm
- Dell PowerConnect M6220 Documentation http://support.dell.com/support/edocs/network/PCM6220/en/index.htm
- Dell PowerConnect M6348 Documentation http://support.dell.com/support/edocs/NETWORK/PCM6348/en/index.htm
- SAN Design Best Practices for the Dell PowerEdge M1000e Blade Enclosure and EqualLogic PS Series Storage (1GbE) http://en.community.dell.com/techcenter/storage/w/wiki/3893.san-design-best-practices-for-the-dell-poweredge-m1000e-blade-enclosure-and-equallogic-ps-series-storage-1gbe-by-sis.aspx
- Best Practices for Enhancing Microsoft Exchange Server 2010 Data Protection and Availability using Dell EqualLogic Snapshots http://en.community.dell.com/techcenter/storage/w/wiki/2633.enhancing-microsoft-exchange-server-2010-data-protection-and-availability-with-equallogic-snapshots-by-sis.aspx
- Sizing Microsoft Exchange 2010 on EqualLogic PS6100 and PS4100 Series Arrays with VMware vSphere 5 http://en.community.dell.com/techcenter/storage/w/wiki/3626.sizing-microsoft-exchange-2010-on-equallogic-ps6100-and-ps4100-series-arrays-with-vmware-vsphere-5-by-sis.aspx

Referenced or recommended Microsoft publications :

- Understanding Exchange 2010 Virtualization http://technet.microsoft.com/en-us/library/jj126252%28EXCHG.141%29.aspx

For EqualLogic best practices white papers, reference architectures, and sizing guidelines for enterprise applications and SANs, refer to Storage Infrastructure and Solutions Team Publications at:

- http://dell.to/sM4hJT