EMC SOFTWARE-DEFINED STORAGE FOR MICROSOFT BUSINESS APPLICATIONS

Storage Sizing Principles and Best Practices

- EMC ScaleIO
- Microsoft Exchange Server
- Microsoft SharePoint Server
- Microsoft Skype for Business

EMC Solutions

Abstract
This white paper provides design guidelines and best practices for Microsoft Exchange, SharePoint and Skype for Business Servers on EMC® ScaleIO®, a software-defined storage solution.

March 2016
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive Summary</td>
<td>4</td>
</tr>
<tr>
<td>Solution design process</td>
<td>6</td>
</tr>
<tr>
<td>Step 1: Application requirements and sizing</td>
<td>8</td>
</tr>
<tr>
<td>Step 2: ScaleIO design considerations for Microsoft business applications</td>
<td>11</td>
</tr>
<tr>
<td>Step 3: Sizing ScaleIO to support Microsoft business applications</td>
<td>22</td>
</tr>
<tr>
<td>Step 4: Implementing and validating Microsoft business applications on ScaleIO</td>
<td>25</td>
</tr>
<tr>
<td>Step 5: Validated test results</td>
<td>34</td>
</tr>
<tr>
<td>Conclusion</td>
<td>38</td>
</tr>
<tr>
<td>References</td>
<td>39</td>
</tr>
</tbody>
</table>
Executive Summary

Business case

Messaging and collaboration tools drive productivity. They are the backbone of a business’s ability to effectively communicate and share ideas and content among employees, partners, and customers. An effective messaging and collaboration suite helps businesses to span geographic divides, and typically consists of business email, instant messaging, contact management, calendars, file sharing, portals, and other social media tools.

Messaging and collaboration tools are used by everyone in the business, from the CEO to individual contributors, enabling them to be more efficient and productive. Online (intranet- or internet-based) collaboration tools enable the rapid exchange of ideas, files, and business assets in an agile manner that is far superior to traditional methods of sharing.

EMC ScaleIO® software offers a high-performance, scalable, and cost-effective storage solution for Microsoft messaging and collaboration applications. With its flexible deployment options, ScaleIO can be installed on servers dedicated to storage services or together with application servers in a hyperconverged configuration.

With ScaleIO, organizations have much greater flexibility. They can start with a small implementation and grow as the business requires. ScaleIO eliminates the need for a large up-front investment in a traditional deployment. ScaleIO offers, instead, a cost-effective and scalable data center-hosted alternative to infrastructure as a service (IaaS) and platform as a service (PaaS) options for most business application use cases.

For customers who want to transform their IT infrastructure to include software-defined storage, knowing how to choose and size components is important. Bringing together messaging and collaboration applications with a software storage solution ensures that customers can meet their short-term collaboration needs while also taking into account their future needs.

Solution overview

This white paper provides high-level guidelines for designing and sizing Microsoft messaging and collaboration applications using a software-defined storage deployment for a specific workload profile. The example storage solution uses Windows Hyper-V for server hardware virtualization and ScaleIO for storage virtualization. ScaleIO protection domain design and the native replication capabilities offered by Microsoft Exchange and SQL Server are used to achieve a highly available solution.

The example solution design uses ScaleIO 1.32 software-defined storage for the following Microsoft applications:

- Exchange Server 2016 for messaging
- SharePoint Server 2013 for collaboration
- Windows Server 2012 R2 Hyper-V as the server hypervisor
- Windows Server 2012 R2 for virtual machines operating system
Figure 1 shows the solution architecture.

**Solution architecture**

**Document purpose**
This white paper provides high-level guidelines for designing and sizing Microsoft messaging and collaboration applications using cost-effective, software-defined, server-based storage with ScaleIO.

The paper also provides ScaleIO deployment best practices for Microsoft Exchange, SharePoint, Skype for Business, and Office Web Apps using a sample deployment scenario and validated test results.

**Audience**
This white paper is intended for EMC partners and customers who have responsibility for the storage design and implementation of the ScaleIO solution, and for Microsoft Exchange, SharePoint, Skype for Business, Virtualization, and Enterprise/Data Center administrators who are evaluating IaaS and PaaS options for these applications.

**We value your feedback!**
EMC and the authors of this document welcome your feedback on the solution and the solution documentation. Contact EMC.Solution.Feedback@emc.com with your comments.

**Authors:** Yunlong Zhang, Aighne Kearney
Solution design process

Overview

ScaleIO is software-defined storage, with design and sizing requirements that are different from traditional storage array solutions. Elements in traditional storage solutions are more clearly divided, comparatively speaking, while software-defined storage elements are mingled together. For example, the server that contributes compute resources for the applications does not serve storage compute resources. Traditional storage arrays have their own processing units, unlike ScaleIO.

This paper will help you understand the relationship between the most important elements and concepts of ScaleIO design and provide you with high-level sizing principles and best practices.

Figure 2 provides an overview of the solution design.

Figure 2. Software defined storage design and sizing overview

To understand how to design and implement your ScaleIO infrastructure for Microsoft messaging and collaboration applications, refer to the requirements and references in this section.

Note: Be sure to read the articles listed in ScaleIO Basics before you start the design process.

Although designing a ScaleIO solution requires background training, you can accomplish it successfully by following these high-level steps:

1. Gather application requirements, including:
Solution design process

- **CPU and memory requirement for each application.**
  Combine this with the compute resources for ScaleIO servers to yield a total compute resource requirement.

- **I/O characteristics of each application.**
  Attempt to understand the major I/O pattern of all the applications when they are running together. This information is used to:
  - Evaluate the total input/output operations per second (IOPS) number, and then use it to estimate the required number of each disk type.
  - Decide whether to enable and tune the read/write cache size for ScaleIO.
  - Put the application data on SSD pools if the usage usually includes large write operations, or frequent and explosive large amounts of writes.
  - Decide the RAID controller’s policy.

- **High availability features at the application level.**
  It is helpful to understand how many copies of the application data are to be maintained in storage. To place the data copies on multiple physical disks and maintain application continuity, you must build more than one protection domain.

For details, refer to Step 1: Application requirements and sizing.

2. **Gather ScaleIO requirement information.**
   Some ScaleIO design decisions (such as the system topology, the network topology, whether to separate ScaleIO network traffic, and so on) are very important because they affect the future scalability and performance of the solution. Solution designers must understand these concepts and evaluate current and future requirements.

   For example, if you are building a small ScaleIO solution with 15 servers but you plan to add up to 200 to 300 servers in the next two years, a flat network topology would not be suitable.

   For details, refer to Step 2: ScaleIO design considerations for Microsoft business applications.

3. **Calculate a starting point for sizing the solution.**
   After gathering the application requirements and some of the ScaleIO design decisions, you can start to calculate:
   - How many disks are needed
   - How many servers are needed
   - How many switches are needed
   - How many NICs each server should contain

For details, refer to Step 3: Sizing ScaleIO to support Microsoft business applications.
Step 1: Application requirements and sizing

4. **Implement the design and validate the performance.**
   Run combined applications on this deployment to determine if the design and sizing results would satisfy the business requirements.

   For details, refer to Step 4: Implementing and validating Microsoft business applications on ScaleIO.

5. **Scale out as necessary.**
   Workloads change as business needs and usage change. The great advantage of ScaleIO is that it can scale out almost linearly, giving you flexibility with both compute resources and storage performance.

   For details, refer to Step 5: Validated test results

For specific user profile information and sizing suggestions on EMC ScaleIO, contact EMC Consulting or your EMC representative. Be ready to provide the I/O workload profile. The EMC representative will assist you with tools and techniques for sizing your environment.

After you have completed these steps, implement your production environment on ScaleIO using your design and sizing results.

**Step 1: Application requirements and sizing**

**Understanding the application requirements**

This section provides a list of key application-related questions to help gather information for your Microsoft messaging and collaboration applications. The following sections provide information to help answer these questions and find a suitable ScaleIO solution for your business needs.

- What applications do I want to deploy in this environment?
- How many users will use these applications?
- What is the main purpose of each application from an end user perspective?
- Based on the expected usage, how many compute resources does each application require?
- How many server roles are there in each application? What topology should you adopt for each application with the server roles?
- What are the availability requirements for each application?
- What is the I/O pattern of each application?
- What is the capacity requirement of each application?
- What will the maximum combined storage workload look like?

*Note:* The Microsoft Exchange Best Practices white paper listed in References describes the typical Exchange I/O pattern. The Skype for Business server typically has a light I/O workload. Refer to the SharePoint Best Practices white paper for SharePoint search I/O characteristics. In a production setting, SharePoint I/O characteristics can vary drastically according to usage. Understanding the I/O pattern of each application you want to deploy means choosing the write disk or caching strategies carefully, as discussed in Step 2: ScaleIO design considerations for Microsoft business applications.
Use the requirements listed in Table 1 when sizing ScaleIO for your Microsoft messaging and collaboration applications. The following sections of this white paper provide information to help you answer these key application questions.

### Table 1. Information requirements and references

<table>
<thead>
<tr>
<th>Application</th>
<th>Requirement</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchange</td>
<td>CPU estimates for different mailbox profiles</td>
<td><em>Sizing Exchange 2016 Deployments</em></td>
</tr>
<tr>
<td></td>
<td>Memory considerations for different mailbox profiles</td>
<td><em>Sizing Exchange 2013 Deployments</em></td>
</tr>
<tr>
<td></td>
<td>Compute sizing and high availability options</td>
<td>(this article also applies to Exchange 2016) <em>Exchange Best Practices White Paper</em></td>
</tr>
<tr>
<td></td>
<td>Exchange I/O characteristics, IOPS evaluation, storage capacity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High availability</td>
<td></td>
</tr>
<tr>
<td>SharePoint</td>
<td>I/O characteristics, IOPS evaluation, storage capacity, typical usage profiles</td>
<td><em>Microsoft SharePoint Server Best Practices and Design Guidelines for EMC storage</em></td>
</tr>
<tr>
<td></td>
<td>farm topology design, compute resource sizing, and search sizing</td>
<td><em>Create a high availability architecture and strategy for SharePoint 2013</em></td>
</tr>
<tr>
<td></td>
<td>High availability</td>
<td></td>
</tr>
<tr>
<td>Skype for Business 2015</td>
<td>Compute resources</td>
<td><em>Server requirements for Skype for Business Server 2015</em></td>
</tr>
<tr>
<td></td>
<td>Sizing</td>
<td><em>Capacity planning user model usage for Skype for Business Server 2015</em></td>
</tr>
<tr>
<td></td>
<td>Bandwidth</td>
<td><em>Skype for Business Bandwidth Calculator</em></td>
</tr>
<tr>
<td></td>
<td>High availability</td>
<td><em>Plan for high availability and disaster recovery in Skype for Business Server 2015</em></td>
</tr>
</tbody>
</table>

### Office Web Apps requirements

For information about Office Web Apps Server compute resources, refer to [Plan Office Web Apps Server on Microsoft Technet](http://technet.microsoft.com). Based on Microsoft performance tests, the number of users that an Office Web Apps server can support varies with the hardware resources used, as detailed in Table 2.

### Table 2. User numbers on Office Web Server with different compute resources

<table>
<thead>
<tr>
<th>Users</th>
<th>vCPUs</th>
<th>RAM</th>
<th>Hard disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 10,000</td>
<td>8 cores</td>
<td>8 GB</td>
<td>60 GB</td>
</tr>
<tr>
<td>Up to 20,000</td>
<td>16 cores</td>
<td>16 GB</td>
<td>60 GB</td>
</tr>
</tbody>
</table>
Step 1: Application requirements and sizing

Evaluate IOPS workload and I/O characteristics

Office Web Apps Server uses the Web App Open Platform Interface protocol (WOPI) to fetch and manipulate files from SharePoint, Skype, and Exchange. The IOPS and I/O characteristics generated by Office Web Apps are no different to the normal file fetching and manipulation operations performed by the messaging and collaboration applications directly. Figure 3 shows the data flow.

![Data flow between application hosts and Office Web Apps Servers](image)

**Figure 3. Data flow between application hosts and Office Web Apps Servers**

Because the enterprise data is stored in SharePoint, Exchange, and Skype for Business, Office Web Apps Server does not have significant I/O operations.

**High availability**

Build your Office Web Apps Server on at least two virtual machines for redundancy, and put them on different virtualization hosts.

For information on load balancing and configuration for Office Web Apps servers, refer to Office Web Apps 2013 multi servers (NLB) Installation and Deployment for SharePoint 2013 step by step guide on Microsoft TechNet.

Summary

After you have gathered the information listed in Table 1 for each application and combined the results, you will know:

- How many CPU resources are required
- How much memory is required
- How much storage capacity is required
- How many IOPS will be generated by these applications
- Approximate I/O characteristics, such as read/write ratio, skew, and I/O size
- What high availability features to implement at the application level

You are now ready to move to the next step: designing the ScaleIO environment for your Microsoft business applications.
Step 2: ScaleIO design considerations for Microsoft business applications

This section discusses how to gather requirements for designing a ScaleIO solution that supports a Microsoft Exchange, SharePoint, Skype for Business, and Office Web Apps workload.

When designing a ScaleIO environment you must determine:

- How many compute resources you plan to give ScaleIO
- Which ScaleIO topology to choose
- How to design the protection domains, storage pools, and fault sets
- Whether to use cache and, if so, how much cache capacity to allocate
- What ScaleIO network topology to use and how to set network-related configurations

Application and storage designers should work together to determine storage requirements. Table 3 provides a list of questions that will help you clarify your storage needs.

Table 3. Storage requirement and references

<table>
<thead>
<tr>
<th>Questions</th>
<th>Reference in this paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will we use existing servers or new servers?</td>
<td>ScaleIO topology deployment plan</td>
</tr>
<tr>
<td>If we use existing servers:</td>
<td>ScaleIO topology deployment plan</td>
</tr>
<tr>
<td>How many do we plan to use in the ScaleIO configuration?</td>
<td></td>
</tr>
<tr>
<td>How many do we plan to set up as a server in ScaleIO? Do they share racks or power?</td>
<td></td>
</tr>
<tr>
<td>Do we need to set up a fault set for these servers in ScaleIO?</td>
<td></td>
</tr>
<tr>
<td>What ScaleIO configuration do we prefer: two-layer converged, single-layer hyperconverged, or mixed?</td>
<td>ScaleIO topology deployment plan</td>
</tr>
<tr>
<td>Will we use existing network switches or purchase new ones?</td>
<td>Network design</td>
</tr>
<tr>
<td>What network topology do we prefer for the network: leaf-spine or flat?</td>
<td></td>
</tr>
<tr>
<td>How should we plan network redundancy?</td>
<td></td>
</tr>
<tr>
<td>How many 10 GbE ports are available on the switches? The switches should have more 10 GbE ports than all the 10GbE NICs the ScaleIO servers combined (to accommodate future growth).</td>
<td>Network throughput</td>
</tr>
<tr>
<td>Are we extending an existing network to accommodate a small number of nodes?</td>
<td>Network performance considerations</td>
</tr>
</tbody>
</table>
## Step 2: ScaleIO design considerations for Microsoft business applications

<table>
<thead>
<tr>
<th>Questions</th>
<th>Reference in this paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are we separating ScaleIO Data Server (SDS) and ScaleIO Data Client (SDC) traffic?</td>
<td>Inter-node communication network and data path network</td>
</tr>
<tr>
<td>Do our servers contain RAID cards?</td>
<td>Caching in ScaleIO</td>
</tr>
<tr>
<td>How many network interface cards (NICs) do these servers have? The number of NICs on ScaleIO servers must match the high availability requirement and throughput that this server contributes.</td>
<td>NICs of networking</td>
</tr>
</tbody>
</table>

The following sections provide information to help you answer these questions.

### ScaleIO compute resources

The following ScaleIO components consume, on average, 5 to 10 percent of these CPU resources: Meta Data Manager (MDM), Tie Breaker (TB), and SDS. Consider the following resource usage when planning the number of servers:

- If an all-flash configuration is pushed to its limit, the cost might be 20 percent for a modern server.
- We do not reserve memory or CPU for ScaleIO because doing so consumes resources. For planning purposes, four cores on a 24-core box use 16 percent of CPU resources, which is more than enough, unless an all-flash configuration is regularly pushed to the upper end of its capabilities.
- 4 GB of memory is more than enough RAM for ScaleIO, unless RAM cache is configured.

Figure 4 shows the data flow between the SDC and SDS components in ScaleIO. For more information, refer to the articles listed in Essential reading.
Step 2: ScaleIO design considerations for Microsoft business applications

After you have determined the total number of ScaleIO servers, you must consider the compute resources consumed by ScaleIO and the applications combined to arrive at the total compute resources requirement. Step 3: Sizing ScaleIO to support Microsoft business applications discusses the number of servers.

ScaleIO implementations have three standard configurations, each of which provides flexibility and scalability.

Figure 5 shows a single-layer hyperconverged infrastructure.

Figure 6 shows a two-layer converged infrastructure.

Figure 7 shows a mixed configuration.
Step 2: ScaleIO design considerations for Microsoft business applications

Choosing a configuration

Before choosing among these configurations, review the descriptions in Table 4.

Table 4. Comparing ScaleIO configurations

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Description</th>
<th>Advantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-layer hyperconverged</td>
<td>Brings the storage, compute, and application resources together in a single layer.</td>
<td>Managing the infrastructure is much simpler. We chose this configuration for our example solution. EMC recommends hyper-converged as the software-defined storage for an IaaS infrastructure.</td>
</tr>
<tr>
<td>Two-layer converged</td>
<td>Enables the customer to move to a software-defined scale-out SAN infrastructure using commodity hardware.</td>
<td>Customers can keep running the applications on separate servers and transform the way that they handle SAN storage. ScaleIO can be configured as a two-layer SAN solution used to connect to an existing application environment.</td>
</tr>
</tbody>
</table>

The choice of configuration affects the compute resource requirements of the solution as a whole. A two-layer or mixed infrastructure requires more compute resources because the compute resource supporting the storage will not be used to support applications.

From an operational efficiency perspective, two-layer is better when multiple teams are managing the environment. Hyperconverged is better when a cross-functional team is managing the environment.
Step 2: ScaleIO design considerations for Microsoft business applications

ScaleIO gives you the flexibility to use any of these configurations. For our example deployment, we chose the one-layer hyperconverged configuration to fully leverage the compute resources of all the servers and make it possible to live-migrate application virtual machines to any of the servers for maintenance purposes.

Using protection domains, storage pools, and fault sets

Protection domain, storage pool, and fault set are important ScaleIO configuration concepts. When designing in a production environment, consider doing the following:

- Use multiple domains to separate performance and prevent storage service interruption caused by multiple failures in a ScaleIO implementation.
- For applications that have high availability options, put data copies on separate protection domains to eliminate a single point of storage failure.
- When a number of SDSs are likely to fail simultaneously (for example, because they are sharing a rack or cooling system), put these SDSs in the same fault set.
- Put HDDs and SSDs in their own storage pools.

Note that performance improves as you add more nodes and disks to a storage pool.

For Microsoft messaging and collaboration applications solutions, our high availability requirement includes Exchange DAG and SQL AlwaysOn for SharePoint databases. Because we require two copies of the data for every application, we configured two protection domains.

For demonstration purposes, we did not set up fault sets in our example solution below.

Modern storage design requires one large pool instead of many small pools for better management and maintenance. We followed best practice in setting up one large HDD pool for our application workloads.

Disks in the ScaleIO server

The maximum number of devices is 64 per server. Any kind of disk (HDD, SSD, or PCI-E card) can be contained in each ScaleIO server.

You can achieve better parallelism and performance by distributing a fixed number of disks across more servers. Because Microsoft messaging and collaboration applications do not have many large write I/Os, a large HDD pool can handle those workloads well. This solution does not use SSDs because they are not required for the workload.

If you have TPC-E workloads on SQL Server, EMC suggests you put your data in an SSD pool or enable the ScaleIO write cache. The choice depends on your I/O characteristics. For example, a large write I/O with a high skew number can benefit more from the write cache. The following section of this document discusses ScaleIO caching.

Caching in ScaleIO nodes

Cache is a critical aspect of storage performance. The ScaleIO cache keeps recently accessed data readily available. I/Os read from the cache have a lower response time than I/Os serviced by the drives, including flash drives.
Read cache

There are two types of read cache: **RAM cache** and **XtremCache**. Both these cache types are managed by the SDS, not the SDC.

- RAM cache is the RAM that is reserved for cache storage devices in the storage pool.
- XtremCache is a software layer located under the SDS. It allows use of any Flash drive type of any size as additional system read cache.

**Note:** Use a read cache for HDD configurations, not for flash-only configurations.

**ScaleI0 1.32:** The default RAM cache size is 128 MB. The supported maximum cache size is 128 GB.

**ScaleI0 2.0:** The maximum size is 300 GB.

Additional cache size can help improve the ScaleIO performance, depending on the application's I/O profile (skew value).

Figure 8 shows the normal I/O flow in ScaleIO.

Figure 8. I/O flow in ScaleIO

Figure 9 shows the cache in ScaleIO architecture.
Figure 9. Read cache in ScaleIO

For more information, refer to *EMC ScaleIO Basic Architecture*.

In our solution, for demonstration proposes, we used the default value for the read cache: 128 MB per SDS node. You can increase this value in your environment if you have enough compute resources.

**Write cache**

In ScaleIO, the RAID controller manages write caching. Writes can also be held in cache to satisfy read requests after they are sent to disk.

RAID controllers (such as LSI and PMC) have battery backup for write buffering. Some RAID controllers also have an option to increase their cache capability to flash drives configured in the systems. This enables you to increase cache from a range of 1 to 4 GB of DRAM to a range of 512 GB to 2 TB if the flash cache is managed by the RAID Controller. This cache is used for both reads and writes.

The main effect of write buffering is that the write response time is much lower when the I/O is acknowledged from DRAM/flash rather than an HDD drive.

An added benefit of buffering the writes in a RAID controller is that it increases the maximum IOPS of an HDD drive. It even reduces the drive's load by eliminating rewrites to the same address locations.

If your application requires a very high IOPS workload and is heavily dependent on large write I/O operations, EMC suggests you enable the write cache for data stored in HDD pools in ScaleIO. Another option is to put the data on ScaleIO SSD pools. The SSD used as write cache can benefit all the data stored in the ScaleIO node. Choose the best method for your solution.

In our example deployment, we did not enable the “Flash as a write cache” feature. Instead, we used the default write buffering provided by the RAID controller.
Step 2: ScaleIO design considerations for Microsoft business applications

**RAID controller policy in ScaleIO nodes**

This section discusses RAID controller configuration and the read and write policies that can affect ScaleIO performance.

**Read policy**

The HDD disk and SSD disk (the normal ScaleIO disk that is not used as a read or write cache) should have a different read policy on the RAID controller. There is minimal pre-fetching as part of the ScaleIO cache code—for example, a sequential read of 512 Byte brings 4 KB into cache.

A best practice recommendation is to use the read-ahead feature in the cache controller only for HDD drives. This enables pre-fetching of I/Os to increase the performance when using HDD drives.

With Flash drives, EMC does not recommend this feature because data is typically placed randomly on SSD media.

**Write policy**

Write policy should be configured differently on HDD and SSD disks, as follows:

- For HDD disks, the write policy should be Write back (or Write back good BBU).
- For flash-only configurations, a pass-through IOC is usually recommended instead of a RAID-On-Chip (ROC) controller because writes are acknowledged from the flash drives directly.

These read/write policy best practices apply not only for Microsoft messaging and collaboration applications but also for other applications workloads on ScaleIO.

**Network design**

ScaleIO deployments can have one of two network topologies: leaf-spine network or flat network.

**Leaf-spine network topology**

A leaf-spine topology is a two-tier architecture consisting of leaf switches and spine switches, as shown in Figure 10.

EMC recommends using a leaf-spine topology for a ScaleIO implementation in most cases. A leaf-spine network provides the following benefits:

- Ability to scale upwards to hundreds of nodes
- Facilitates scale-out deployments without the need to re-architect the network (future-proofing)
- Non-blocking when designed correctly to allow for maximum bandwidth
- All connections have equal access to bandwidth
- Predictable latency
- Highest availability and performance
Flat network topology

A flat network design is easier and less costly to implement, and may be preferred if you are extending an existing network or if the network does not scale beyond four switches, as in Figure 11. A flat network is the fastest and simplest way to get your ScaleIO deployment up and running.

The primary use cases for a flat network topology are:

- Small deployment, not extending beyond four switches
- Remote office/back office (ROBO)
- Small business

Internode communication network and data path network

In ScaleIO, internode communication can occur in one of two ways, as follows:

- Over a separate IP network, with access to the other ScaleIO components
- On a single IP network
Step 2: ScaleIO design considerations for Microsoft business applications

For more information about the available combinations, refer to the *EMC ScaleIO 1.32 User Guide*. Base your choice on your organization's requirements, security considerations, performance needs, and IT environment. Table 5 describes the characteristics you should take into account.

**Table 5. Network type considerations**

<table>
<thead>
<tr>
<th>Network type</th>
<th>Characteristics</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single IP network</td>
<td>Ease of use, fewer IP addresses required.</td>
<td></td>
</tr>
<tr>
<td>Multiple separate IP networks</td>
<td>Security, redundancy, performance, separate customer data and ScaleIO internal management data.</td>
<td>Control network—Used for MDM control communications with SDSs and SDCs. Also used to convey data migration decisions, although no user data passes through the MDM. Rebuild and rebalance network—Used for SDS-SDS communication only. These IP addresses are used for rebuild and rebalance operations and also to communicate with the MDM. Data path network—Used for SDS-SDC communication. These addresses are used for read/write user data operations.</td>
</tr>
</tbody>
</table>

Network separation guarantees network bandwidth, which means a much faster rebuild process for ScaleIO and a lower risk of storage outage. Under separated networks, the applications do not have to share bandwidth with any ScaleIO rebuilding or rebalancing process. Separated networks are more costly, consuming more switch ports and 10 GbE NICs. Separating the networks is a good choice for business-critical usage of Microsoft messaging and collaboration applications. In our solution, we did not separate the network traffic because our I/O pattern research indicated that throughput is not high.

This section discusses performance features to consider when designing a network.

**Throughput**

The throughput of the ScaleIO system comes from the contribution of each SDS node. When SDS nodes have enough disks, they might be all running simultaneously and
consequently increase the total bandwidth. The network could become a bottleneck, especially during rebuilding and rebalancing operations.

When ScaleIO is deployed on a 1 Gbps network, storage performance is typically bottlenecked by the network capacity. EMC recommends using 10 Gbps network technology at a minimum.

**Latency**

Network latency is an important consideration when designing your network. Minimizing the amount of network latency provides for improved performance. For best performance, the latency for all SDS and MDM communication should not exceed a roundtrip time of 1ms.

You can verify this by using a ping command between the SDS and the MDM, or more comprehensively by using the SDS Network Latency Meter Test tool. Refer to *EMC ScaleIO Networking Best Practices* for more information about this tool.

ScaleIO is not designed to extend outside the data center. EMC discourages the use of wide area networks to operate ScaleIO.

**NIC configuration best practices**

When completing your NIC configuration, consider the following:

- Although ScaleIO supports single and multiple NIC configurations, as a best practice EMC recommends at least two 10 Gbps Ethernet NICs per ScaleIO node as an initial configuration.
- As a baseline for system design, every ScaleIO SDS node should contain at least two network interfaces for redundancy.
- Additional network capacity may be required for a variety of reasons. ScaleIO lets you scale network resources by adding network interfaces.
- Although not required, there may be situations where isolating front-end and back-end traffic for the storage network may be ideal. In all cases, we recommend a minimum of two NICs for redundancy, capacity, and speed. The primary motive for segmenting front-end and back-end network traffic is to guarantee the performance of storage-related and application-related network traffic.

Other network-related techniques may also affect performance, such as NIC pooling, dedicated access ports, jumbo frames flow control, link aggregation, and Infiniband. For more information, refer to the *EMC ScaleIO Networking Best Practices* white paper.

Next, size these components:

- ScaleIO topology
- Network topology
- Network traffic separation
- Protection domains, fault sets, and pools
- Cache
Step 3: Sizing ScaleIO to support Microsoft business applications

Overview

After completing the application design and the ScaleIO design, the next step is to size the solution. This involves determining the number of disks and servers you need.

Sizing guidelines

When sizing a ScaleIO solution, observe the following best practices:

- Consider the sizing application requirement first, and then choose the infrastructure that suits the requirement. This is known as the two-phase approach.

- The compute resources, networking resources, and storage resources should be sufficient for the expected infrastructure requirement and for future growth. These resources should match each other in the solution:
  - Add up the compute resources of all the applications and ScaleIO nodes to obtain the total required CPU and memory.
  - Network resources, including:
    - High bandwidth, deep buffer switches for spine network
    - Available 10 Gbps switch port numbers
    - 10 Gbps NICs on each ScaleIO node
  - Storage resources, including:
    - Number and type of disk media types (HDD, SSD, PCI-E card)
    - Number of ScaleIO servers
    - Memory assigned to ScaleIO as read cache
    - RAID controller hardware on hosts
    - RAID controller software to configure flash disk to increase cache capability (such as LSI CacheCade and PMC MaxCache)
    - When sizing ScaleIO storage, consider both capacity and performance and choose whichever one requires a greater number of disks—for example, if you need 100 disks to meet the performance requirement and up to 200 disks to meet the capacity requirement, deploy 200 disks.

- When sizing the number of ScaleIO nodes, consider both CPU and memory and choose whichever one requires a greater number of servers. Notice that different types of ScaleIO topology also affect the compute resources required.

- ScaleIO scales almost linearly with regard to capacity and performance.

- Size from both the capacity perspective and performance perspective, then choose the one that requires more disks and servers as a starting point. After validation, add more servers or disks if necessary.
Sizing for capacity

ScaleIO reserves space on servers for an unplanned outage when rebuilds are going to require unused disk space. To ensure data protection during server failures, ScaleIO reserves a certain amount of the capacity by default, and does not allow it to be used for volume allocation. How much spare capacity ScaleIO reserves by default depends on the number of nodes. For example, with three nodes it is 33 percent; with 10 nodes, it is 10 percent; and with 20 nodes, it is 5 percent.

For full system protection in the event of a node failure, ensure that the spare capacity is at least equal to the amount of capacity in the node containing the maximum capacity or the maximum fault set capacity. If all nodes have equal capacity, we recommend setting the spare capacity value to at least 1/N of the total capacity, where N is the number of SDS nodes.

To estimate capacity in the ScaleIO environment, use the following calculation:

\[
\text{Estimated available capacity} = \left( \frac{\text{Total raw capacity of all disks}}{2} \right) \times (1 - \text{percentage of reservation})
\]

For example, if a ScaleIO environment has only eight nodes, then 1/8 capacity is reserved to ensure data protection during a server failure. If a ScaleIO environment has specified three fault sets, the available capacity of the other two fault sets must be able to handle all the data from one failed fault set.

Sizing for storage performance

When sizing ScaleIO performance, keep in mind that:

- ScaleIO scales almost linearly.
- The value of IOPS per disk in ScaleIO systems varies on different I/O sizes, read/write ratios, and disk types.
- The network bandwidth should satisfy the total disk bandwidth on each host. Refer to Sizing for NICs on ScaleIO servers for more information.
- MDM traffic and SDS-SDS traffic should always be on 10 Gbps networks with redundancy, whether they are single or separated networks.
- To deal with intense write workloads, in addition to using a pool of Flash drives or scaling out the number of nodes and disks, consider enabling Flash drives as a write cache on RAID controller. Refer to Caching in ScaleIO nodes for more information.
- The more disks and nodes a ScaleIO pool has, the better the performance. Performance here means the application’s storage performance and also the ScaleIO rebuild and rebalance performance. A faster rebuild means a lower chance of another disk failure during the rebuilding process, bringing more security to the storage. In addition, one large pool can make management easier than many small pools.

Sizing for NICs on ScaleIO servers

To optimize throughput and thus minimize the ScaleIO rebuilding and rebalancing time, use the following calculation:

\[
\text{Number of drives per server} \times \text{average sequential drive bandwidth in MB/s} \quad \text{is approximately equal to} \quad \text{number of 10G NICs} \times 10,000 / 8
\]
Step 3: Sizing ScaleIO to support Microsoft business applications

For guidance, here are some typical values:

- An HDD drive can achieve up to 100 MB/s of application reads and up to 50 MB/s of application writes.
- An SSD drive can achieve about 450 MB/s of application reads and up to 50 MB/s of application writes.
- A 1 GbE NIC can achieve roughly 100 MB/s of application reads and up to 50 MB/s of application writes.
- A 10 G NIC can achieve roughly 1,000 MB/s of application reads and up to 500 MB/s of application writes.

For a typical ScaleIO SDS containing 10 HDD drives, use the following calculation:

$$10 \text{ drives} \times 100 \text{ MB/s} = 1,000 \text{ MB/s maximum bandwidth}$$

One 10 G NIC is sufficient for normal usage, but two 10 G NICs are required for high availability.

For an SDS containing six SSD drives, use the following calculation:

$$6 \text{ drives} \times 450 \text{ MB/s} = 2,700 \text{ MB/s maximum bandwidth}$$

In this example, three 10 GbE NICs are required for redundancy and best performance.

Summary

The sizing guidelines and best practices introduced in this section give you a good starting point for designing your solution. If the actual workload is higher than anticipated or grows beyond expectation, it can be easily scaled out by adding more nodes and disks to the ScaleIO solution.
Step 4: Implementing and validating Microsoft business applications on ScaleIO

Overview
This section describes an example ScaleIO solution supporting an application profile that consists of up to 1,500 users using the following applications:

- SharePoint 2013
- Exchange 2016

ScaleIO server configuration
In our tests, we used Windows Server 2012 R2 as the hypervisor and platform for ScaleIO. All the virtual machines run on Windows Server 2012 R2 failover cluster.

We installed the Windows operating system (OS) of the ScaleIO server on a RAID 1 virtual disk provided by the RAID controller. This was to prevent the single point of failure for the disk having the ScaleIO server OS.

Workload profile
A small company has 1,500 users. They want to set up a ScaleIO solution for their Microsoft business applications, including Exchange and SharePoint. A large mailbox is a must for Exchange (for capacity) and the SharePoint usage is focused on document management (more I/O). The business has Skype for Business and Office Web Apps Server, but wants to build and validate the solution on Exchange and SharePoint first.

Exchange profile
The user load is heavy (10 GB mailbox, 150 messages/user/day) in this sample deployment scenario. Table 6 details the Exchange Server 2016 workload profile.

Table 6. Exchange Server 2016 workload profile

<table>
<thead>
<tr>
<th>Profile characteristic</th>
<th>Quantity/Type/Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Exchange Server 2013 users</td>
<td>1,500</td>
</tr>
<tr>
<td>Number of Mailbox Server virtual machines</td>
<td>2</td>
</tr>
<tr>
<td>Number of DAGs and database copies</td>
<td>1 DAG with 2 database copies</td>
</tr>
<tr>
<td>Number of users per Mailbox Server</td>
<td>1,500 total mailboxes</td>
</tr>
<tr>
<td></td>
<td>(750 active and 750 passive during normal operating conditions)</td>
</tr>
<tr>
<td>User profile (in DAG configuration)</td>
<td>150 messages/user/day (0.101 IOPS)</td>
</tr>
<tr>
<td>Read/write ratio</td>
<td>3:2 in a DAG configuration</td>
</tr>
<tr>
<td>Mailbox size</td>
<td>10 GB</td>
</tr>
<tr>
<td>24 x 7 background database maintenance configuration</td>
<td>Enabled</td>
</tr>
</tbody>
</table>

SharePoint profile
This is a Document Management Portal usage scenario. Table 7 details the SharePoint Server 2013 workload profile.
Step 4: Implementing and validating Microsoft business applications on ScaleIO

Table 7. SharePoint Server 2013 workload profile

<table>
<thead>
<tr>
<th>Profile characteristic</th>
<th>Quantity/Type/Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total user count</td>
<td>1,500 users</td>
</tr>
<tr>
<td>Usage profiles</td>
<td>50%/20%/20%/10%</td>
</tr>
<tr>
<td>(browse/search/modify/upload)</td>
<td></td>
</tr>
<tr>
<td>User concurrency rate</td>
<td>50%</td>
</tr>
<tr>
<td>Content database sizes</td>
<td>1.5 TB</td>
</tr>
<tr>
<td>Number of content databases</td>
<td>3</td>
</tr>
<tr>
<td>Document size range</td>
<td>200 KB to 2 MB. On average 335 KB</td>
</tr>
<tr>
<td>Total site collection count</td>
<td>7</td>
</tr>
<tr>
<td>Size per site</td>
<td>10 GB to 400 GB</td>
</tr>
<tr>
<td>Total site count</td>
<td>70</td>
</tr>
</tbody>
</table>

Table 8. Requirements for our deployment example

<table>
<thead>
<tr>
<th>Question</th>
<th>Requirements for our sample deployment scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>What applications do I want to deploy in this environment?</td>
<td>• SharePoint 2013</td>
</tr>
<tr>
<td></td>
<td>• Exchange 2016</td>
</tr>
<tr>
<td></td>
<td>• Skype for Business 2015 (designed but not tested in our validation)</td>
</tr>
<tr>
<td></td>
<td>• Office Web Apps (designed but not tested in our validation)</td>
</tr>
<tr>
<td>How many users will use these applications?</td>
<td>About 1,500 users</td>
</tr>
<tr>
<td>What is the main purpose of each application from an end user perspective?</td>
<td>• SharePoint is used mainly for document management.</td>
</tr>
<tr>
<td></td>
<td>• Exchange has heavy usage.</td>
</tr>
<tr>
<td></td>
<td>• Skype for Business and Office Web Apps have normal usage.</td>
</tr>
<tr>
<td>Based on the usage of each application, how many compute resources do we need?</td>
<td>Six Cisco C240 M3 servers will be sufficient for the compute resources requirement.</td>
</tr>
<tr>
<td></td>
<td>If storage requires more than six servers, we will use the higher number in sizing.</td>
</tr>
</tbody>
</table>
Step 4: Implementing and validating Microsoft business applications on ScaleIO

### Question

How many server roles are there in each application? What topology will we adopt for each application with the server roles?

- For Exchange 2016, we used Jetstress to test the storage performance.
- For SharePoint:
  - Two web servers
  - Four application servers
  - Two database servers

### Requirements for our sample deployment scenario

What high availability features of each application do we want to adopt?

Refer to [high availability feature design](#) for more information.

What is the I/O pattern of each application?

Refer to [Application characteristics and considerations on ScaleIO](#).

What is the capacity requirement of each application?

- 7 TB SharePoint
- 30 TB for Exchange
- 3 TB for VMOS

What is the I/O pattern of each application?

Generally, it is random I/O. Read is greater than write. The I/O workloads are not heavily write-intensive.

Table 9 lists our answers about storage requirements for the applications.

### Table 9. Storage requirements for this solution deployment example

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Test deployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will we use existing servers or new servers?</td>
<td>We will use existing servers.</td>
</tr>
<tr>
<td>How many existing servers do we plan to use in the ScaleIO configuration?</td>
<td>We have 15 Cisco C240 servers in total. All of them can be used.</td>
</tr>
<tr>
<td>If we use the existing servers, how many do we plan to set up as a server in ScaleIO?</td>
<td>For demonstration purposes, we did not configure fault sets.</td>
</tr>
<tr>
<td>Do the existing servers share racks or power?</td>
<td></td>
</tr>
<tr>
<td>Do we need to set up fault sets for these servers in ScaleIO?</td>
<td></td>
</tr>
<tr>
<td>If we use the existing servers, how many do we plan to set up as a node in ScaleIO?</td>
<td>We have 15 servers and we can use them all.</td>
</tr>
<tr>
<td>Where are the existing servers located?</td>
<td>Scattered around the data center, not sharing a shelf or power. Accordingly, no fault set is needed.</td>
</tr>
<tr>
<td>What ScaleIO configuration do we want to choose: two-layer, hyper-converged, or mixed?</td>
<td>One-layer hyperconverged infrastructure.</td>
</tr>
<tr>
<td>Will we use existing network switches or purchase new ones?</td>
<td>Existing network switches. The 10 G ports are sufficient.</td>
</tr>
<tr>
<td>Requirement</td>
<td>Test deployment</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>What network topology are we going to use for the network: leaf-spine or flat?</td>
<td>This is a validation solution. We used existing switches and we do not plan to scale out the infrastructure in the future. A flat network topology is more suitable.</td>
</tr>
<tr>
<td>How do we plan the network redundancy?</td>
<td>Redundant switches</td>
</tr>
<tr>
<td>How many 10 GbE ports are available?</td>
<td>We have more than 30 ports available to connect all the ScaleIO servers.</td>
</tr>
<tr>
<td>Are we extending an existing network to accommodate a small number of nodes?</td>
<td>We will use our existing devices to build a new flat network topology for our ScaleIO solution.</td>
</tr>
<tr>
<td>Are we separating SDS and SDC traffic?</td>
<td>• Not in this solution.</td>
</tr>
<tr>
<td></td>
<td>• SDS-SDS network is used for rebuild and rebalance operations and for communication with the MDM.</td>
</tr>
<tr>
<td></td>
<td>• SDS-SDC network is used for read/write user data operations.</td>
</tr>
<tr>
<td></td>
<td>• Separating the network eliminates the network conflict between the two kinds of traffic.</td>
</tr>
<tr>
<td></td>
<td>• Combining the information relating to disk throughput and application workload, and based on the calculation in Throughput, separation is not necessary.</td>
</tr>
<tr>
<td>Do our servers contain RAID cards?</td>
<td>Yes, all of them. We followed the best practices described in Caching in ScaleIO to set up the read and write policy in the RAID controller.</td>
</tr>
<tr>
<td>How many network interface cards (NICs) do these servers have?</td>
<td>• Each server has two 10 GbE NICs for ScaleIO network redundancy.</td>
</tr>
<tr>
<td></td>
<td>• Each server has two 1 GbE NICs for application network redundancy, that is, each server has at least two 10 GbE NICs, and two 1 GbE NICs.</td>
</tr>
<tr>
<td></td>
<td>• Combining the disk throughput and application workload information and based on the calculation in Throughput, two 10 GbE NICs are sufficient to handle the application workload and the possible rebuild and rebalance workload without affecting applications.</td>
</tr>
</tbody>
</table>

Table 10 details the hardware resources used in the solution.
Table 10. Hardware resources

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Quantity</th>
<th>Detail</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Server</td>
<td>15</td>
<td>CPUs: 2<em>16-core 2.60 GHz Memory: 128 GB Disks: 10</em>900 GB RAID controller: LSI MegaRAID SAS</td>
<td>Hyper-V host as ScaleIO server</td>
</tr>
<tr>
<td>Switch</td>
<td>4</td>
<td>• 10 GbE * 2</td>
<td>• Two for internal host connection using 10 GbE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 1 GbE * 2</td>
<td>• Two for external management connection using 1 GbE</td>
</tr>
</tbody>
</table>

Table 11 details the software resources used in this solution.

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMC software</td>
<td>ScaleIO 1.32</td>
</tr>
<tr>
<td>Microsoft software</td>
<td></td>
</tr>
<tr>
<td>• Jetstress 2013 (Exchange 2016 DLLs)</td>
<td></td>
</tr>
<tr>
<td>• Windows Server 2012 R2</td>
<td></td>
</tr>
<tr>
<td>• Hypervisor</td>
<td></td>
</tr>
<tr>
<td>• Operating system for VMs</td>
<td></td>
</tr>
<tr>
<td>• SharePoint 2013 SP1 with November 2015 CU</td>
<td></td>
</tr>
<tr>
<td>• SQL Server 2014 (for SharePoint backend)</td>
<td></td>
</tr>
</tbody>
</table>

Given the industry-level HDD disk performance characteristics, 150 HDD 900 GB disks are a good starting point to support validation for our solution.

**Application design**

After analyzing the requirements, we designed the applications as detailed in Table 12.

**Virtual machine design**

Table 12. Virtual machines configuration

<table>
<thead>
<tr>
<th>Virtual machine role</th>
<th>Quantity</th>
<th>vCPU</th>
<th>Memory (GB)</th>
<th>Boot disk VHDX (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharePoint - Web server</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>SharePoint - Application server (with crawler, content processing, admin, and analytics component)</td>
<td>2</td>
<td>12</td>
<td>12</td>
<td>100</td>
</tr>
<tr>
<td>SharePoint - Application server (with query processing and index partition component)</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td>100</td>
</tr>
</tbody>
</table>
Step 4: Implementing and validating Microsoft business applications on ScaleIO

<table>
<thead>
<tr>
<th>Virtual machine role</th>
<th>Quantity</th>
<th>vCPU</th>
<th>Memory (GB)</th>
<th>Boot disk VHDX (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SharePoint - SQL server (AlwaysOn AG enabled)</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>100</td>
</tr>
<tr>
<td>Exchange – Jetstress</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>Office Web Apps</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>Skype for Business – Front end</td>
<td>3</td>
<td>16</td>
<td>32</td>
<td>100</td>
</tr>
<tr>
<td>Skype for Business – Back end</td>
<td>2</td>
<td>16</td>
<td>32</td>
<td>100</td>
</tr>
</tbody>
</table>

**High availability feature design**

Our design included the following high availability features:

- Network load balance for Web frontend servers of SharePoint
- Mirrored search index for SharePoint search
- AlwaysOn Availability Groups for SharePoint databases
- Two copies of DAG for Exchange databases
- Redundant Exchange mailbox servers
- Redundant Office Web App Servers
- Redundant Skype for Business front-end servers
- AlwaysOn Availability Groups for Skype for Business databases

We set anti-affinity rules on the virtual machines in the redundant application server role. Refer to the TechNet article [AntiAffinityClassNames](#) for more information.

**Hyper-V cluster design and configurations**

This solution deploys a Hyper-V cluster consisting of 15 nodes. The cluster configures a file share quorum witness, and all nodes have a vote.

For all virtual machines, we applied the default setting of prevent failback. Therefore, the virtual machines will not automatically move back to the original node. This is because failback uses Quick Migration (the cluster shuts down the virtual machine first before moving it) but not Live Migration.

All the virtual disks used by the virtual machines are configured as dynamically expanding VHDX in this solution.

**Storage design**

**Topology**

We chose a hyperconverged configuration for our deployment because we wanted the flexibility to migrate the virtual machines among all the hosts. Following the best practices in Sizing Guidance, we used all 15 of our existing servers.

**Networking**

We chose a flat network topology for our solution. Using the method described in network throughput, we know the network will not be a bottleneck. Figure 12 shows the converged topology and the networking topology.
Step 4: Implementing and validating Microsoft business applications on ScaleIO

Protection domain, storage pool, and fault set

Two copies of each application's data are needed to fulfill our high availability requirements. To protect the data and prevent a single point of failure in storage, we designed the storage on two protection domains. Each protection domain has one copy of the application data.

We set up one large pool in each of the protection domains to achieve better performance and greater ease of management.

We did not set up fault sets in the test deployment for performance validation. In the production environment, customers must set up servers that share resources or might shut down together in the same fault set.

Because our workload was not compute-intensive or write-intensive, we used only HDD disks in our storage pool. Figure 13 shows the pool configuration. For an SQL Server workload, or a higher customized storage usage SharePoint requirement, EMC recommends putting the data on SSD pools.
Step 4: Implementing and validating Microsoft business applications on ScaleIO

**Application data layout**

We put every application in its own volume. We could therefore set throttling on the volumes in ScaleIO in the event of sudden workload bursts in any application to ensure that other applications would not be affected.

Figure 14 shows the detailed application data layout.

**Figure 13. ScaleIO storage pool configuration**

**Figure 14. Application data layout and volume setup in this ScaleIO solution**
Step 4: Implementing and validating Microsoft business applications on ScaleIO

Caching and RAID controller policy
All our servers have a RAID controller installed. Because our workload is not write-intensive, we did not enable the extended flash drives feature into the RAID controller cache.

- We used the write cache using the RAID controller’s out-of-box write buffering (with battery backup).
- We also used the out-of-box read cache size of 128 MB per SDS. Note that this can be set higher to achieve better performance as necessary.
- Because we only used HDD disks in this solution, the read policy of all disks is read ahead. The write policy of all disks is write back.

This solution ensures that the business can continue in the scenarios described below.

Application-level disturbance
Redundant server roles provide protection against:

- Virtual machine restarts, shutdown, or failure of each application server role:
  - SharePoint Web Servers
  - SharePoint APP server
  - SharePoint database server
  - Exchange mailbox server
  - Skype for business front end
  - Skype for business back end
  - Office Web App Server

AlwaysOn Availability Groups provide protection against:

- Local database corrupted pages in:
  - SharePoint databases
  - Skype for Business databases
- One SQL Server instance outage (automatic failover)

Infrastructure-level disturbance
Because two protection domains are available, application users are not affected if one or more hosts shut down simultaneously in one protection domain. Exchange DAG and AlwaysOn automatic failover ensure business continuity.

Because of the protection domain, application users are not affected if:

- One host in each protection domain restarts or shuts down
- One disk in each protection domain is corrupted

Because of network redundancy, application users are not affected if:

- One 10 GbE switch fails
Step 5: Validated test results

Overview

After designing and implementing the solution, your next step is to test to see if it meets your requirements.

For our example solution, we validated the performance of SharePoint and Exchange running together on our ScaleIO deployment. To simulate a client workload for SharePoint Server, and Exchange Server, we used the following tools:

- Microsoft SharePoint 2013 Visual Studio Team System (VSTS)-generated custom workload
- Microsoft Exchange Jetstress

Combined test results

In our validation test, we ran Exchange and SharePoint together for four hours.

**Note:** Skype for Business and Office Web Apps storage requirements are not as high as for Exchange and SharePoint. These two applications were not included in the validation tests.

SharePoint test result

This result indicates the tested user load on the SharePoint production farm with two Web Front End Servers (WFEs). Table 13 details the test results.

Table 13. SharePoint test results

<table>
<thead>
<tr>
<th>Performance counter</th>
<th>Target</th>
<th>Test result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total passed tests/sec</td>
<td>12.5</td>
<td>17.2</td>
</tr>
<tr>
<td>Maximum user capacity (50% currency)</td>
<td>1,500</td>
<td>2,064</td>
</tr>
<tr>
<td>Browse average test time</td>
<td>&lt;3 sec</td>
<td>0.21</td>
</tr>
<tr>
<td>Search average test time</td>
<td>&lt;3 sec</td>
<td>0.80</td>
</tr>
<tr>
<td>Modify average test time</td>
<td>&lt;3 sec</td>
<td>0.32</td>
</tr>
<tr>
<td>Upload average test time</td>
<td>&lt;3 sec</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Table 14 details the storage performance for the SharePoint database primary replica.

**Table 14. SharePoint primary replica storage performance**

<table>
<thead>
<tr>
<th>Performance counter (SQL primary)</th>
<th>Results (content DB/Log/Service)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Disk sec/Transfer</td>
<td>0.016/0.003/0.010</td>
</tr>
<tr>
<td>Max Disk Transfers/sec</td>
<td>715/343/819</td>
</tr>
<tr>
<td>Disk Transfers/sec</td>
<td>18.57/31.35/19.63</td>
</tr>
</tbody>
</table>

Table 15 details the storage performance of the SharePoint database secondary replica.

**Table 15. SharePoint secondary replica storage performance**

<table>
<thead>
<tr>
<th>Performance counter (SQL Secondary)</th>
<th>Results (content DB/Log/Service)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Disk sec/Transfer</td>
<td>0.004/0.017/0.007</td>
</tr>
<tr>
<td>Max Disk Transfers/sec</td>
<td>928/339/812</td>
</tr>
<tr>
<td>Disk Transfers/sec</td>
<td>18.87/32.89/5.67</td>
</tr>
</tbody>
</table>

The CPU usage of the two Web Servers on the server farm is below 70 percent.

The test results show that the host read latency, write latency, and the overall latency of the virtual disks hosting database primary replicas are all below 20 milliseconds (ms). The search index disk latency values are all below 10 ms.

Each test operation finished within one second, well ahead of the test target of three seconds.

**Exchange test result**

Table 16 shows the average I/O and average latency on the Mailbox server. The performance of the Exchange organization exceeded the design target. In this solution, the limiting factor was the capacity of the Exchange storage pool to support 1,500 users with a 10 GB mailbox size in a two-copy Exchange 2016 DAG configuration. From a performance perspective, the storage pool could potentially support an even greater workload.

**Table 16. Exchange Jetstress test result**

<table>
<thead>
<tr>
<th>Database I/O</th>
<th>Target values</th>
<th>Single Mailbox server</th>
</tr>
</thead>
</table>
| Achieved transactional IOPS (I/O database reads/sec + I/O database writes/sec) | Number of mailboxes * Exchange Server 2013 user IOPS profile  
In this solution: 1500 * 0.101 = 151.5 | 293.6 |
| I/O database reads/sec                                 | N/A (for analysis purpose)                        | 207                   |
| I/O database writes/sec                                | N/A (for analysis purpose)                        | 85.6                  |
| I/O database reads average latency (ms)                | Less than 20 ms                                   | 12.72                 |
### Step 5: Validated test results

<table>
<thead>
<tr>
<th>Database I/O</th>
<th>Target values</th>
<th>Single Mailbox server</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/O log writes average latency (ms)</td>
<td>Less than 10 ms</td>
<td>5.46</td>
</tr>
</tbody>
</table>

In this validation test, the disk subsystem was not pushed to its limit. This sample solution deployment used the number of HDD disks specified to support the capacity. The solution can handle a greater storage workload.

**ScaleIO performance**

Figure 15 shows an example of the virtual SAN’s real-time status captured from the ScaleIO GUI during the combined application test. The ScaleIO system had a total of 1,523 total IOPS. This value can be considered as the average value during the test.

![ScaleIO real time performance screenshot](image15.png)

**Figure 15. ScaleIO real time performance screenshot**

Pool1, for the purpose of testing the worst-case scenario, hosts up to half of all the virtual machines, the SharePoint database primary replica, and all the Exchange Jetstress databases. On Pool1, there are 1,365 IOPS.

On Pool2, which hosts the other half of the virtual machines and the SharePoint database secondary replica, there are 158 IOPS.

The workload was distributed effectively on all devices in the storage pools. During the combined test, the device latencies were almost all under 10 ms, as Figure 16 shows.
Step 5: Validated test results

![Image of device details]

The test results show that the performance of this deployment satisfies the specified requirements.

We did not scale out and revalidate in this example solution and we did not push this ScaleIO deployment to its performance limit. Many compute resources were left available for other applications and future growth. From a storage perspective, the IOPS and latency were not high in this test. The system can handle a greater workload.

In a production environment, the next step would be to implement this design.

**Figure 16. Example of real-time device latency on a ScaleIO node**
Conclusion

Summary

EMC ScaleIO provides a software-only server-based SAN that converges storage and computing resources to form a single-layer, two-layer, or mixed enterprise-grade storage product. ScaleIO storage is elastic and delivers an almost linearly scalable performance. Its scale-out server SAN architecture can grow from just a few to thousands of servers.

This white paper demonstrated ScaleIO design and sizing best practices by using an example of building a production environment with mixed Microsoft messaging and collaboration applications.

The validation test shows that ScaleIO supports Exchange and SharePoint, with the ability to scale out both capacity and performance.

Findings

Table 17 summarizes the key objectives that this solution meets.

<table>
<thead>
<tr>
<th>Objective</th>
<th>Key results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Converged physical infrastructure</td>
<td>ScaleIO uses servers’ direct-attached storage (DAS) and aggregates all disks into global, shared block storage. ScaleIO features single-layer compute and storage architecture without requiring additional hardware or cooling/power/space.</td>
</tr>
<tr>
<td>Simplified design</td>
<td>By integrating Windows Server 2012 R2 Hyper-V with ScaleIO virtual SAN, the hardware resources requirement in this solution is significantly lower than in traditional storage solutions, which leads to a higher server utilization rate and reduced energy and operational costs. In this example solution, the entire environment consists of only 15 physical servers.</td>
</tr>
<tr>
<td>Use the design to support a mixed Microsoft workload on a Windows Server 2012 Hyper-V virtualization platform over ScaleIO virtual SAN and validate the performance level</td>
<td>ScaleIO takes advantage of parallel I/O processing, with all servers participating and sharing I/O loads. The test results show that the designed architecture satisfies all recommended performance guidelines provided by Microsoft for SharePoint Server 2013 and Exchange Server 2016. SharePoint works well on ScaleIO virtual SAN. The designed architecture could easily support 1,500 users with AlwaysOn AG protection configuration. Exchange Server 2016 is successfully deployed on ScaleIO virtual SAN. The designed architecture could easily support 1,500 typical users with 10 GB mailbox size in a failover situation of the two-copy Exchange 2016 DAG configuration, where all workload was handled by a single Exchange server and a single storage pool. The limiting factor in the solution was the available storage capacity. With larger disks, the architecture could support even more users with a similar configuration.</td>
</tr>
</tbody>
</table>
References

Essential reading

EMC recommends that you read the following documents, available from the ScaleIO space in the EMC Community Network or from EMC.com. Access to these documents depends on your login credentials. If you do not have access to a document, contact your EMC representative.

ScaleIO Basics

- *EMC ScaleIO Basic Architecture*
- *EMC ScaleIO Networking Best Practices*
- *Configuring ScaleIO and XtremCache*
- *EMC ScaleIO 1.32 User Guide*

Application sizing and design best practices

- *Microsoft Exchange Server Best Practices and Design Guidelines for EMC Storage*
- *Microsoft SharePoint Server Best Practices and Design Guidelines for EMC Storage*

Microsoft documentation

The following documentation on the Microsoft TechNet website provides additional and relevant information:

- Exchange Server 2016
- SharePoint 2013
- Skype for Business Server 2015
- Windows Server 2012