EMC AVAILABILITY FOR EXTREME PERFORMANCE OF MICROSOFT SQL SERVER

EMC XtremIO, EMC VPLEX, VMware vSphere

- Support very high throughput for OLTP SQL Server workloads
- Virtualize and consolidate database instances
- Continuous availability with EMC VPLEX Metro

EMC Solutions

Abstract

This white paper describes the advantages of highly available virtualized Microsoft SQL Server 2012 and 2014 database failover clusters deployed on an EMC® XtremIO™ all-flash array enabled by EMC VPLEX® Metro.

September 2014
# Table of contents

**Executive summary** .................................................................................................................. 6  
  Business case .............................................................................................................................. 6  
  Solution overview ......................................................................................................................... 6  
  Key results ................................................................................................................................... 7  

**Introduction** .............................................................................................................................. 8  
  Purpose ........................................................................................................................................ 8  
  Scope .......................................................................................................................................... 8  
  Audience ...................................................................................................................................... 8  
  Terminology .................................................................................................................................. 8  

**Technology overview** ............................................................................................................... 9  
  Overview ...................................................................................................................................... 9  
  EMC XtremIO ................................................................................................................................. 9  
    Writeable snapshots ..................................................................................................................... 9  
    XtremIO Management Server .................................................................................................... 10  
  EMC VPLEX ................................................................................................................................. 10  
    EMC VPLEX clustering architecture ......................................................................................... 10  
  VMware vSphere .......................................................................................................................... 11  
  Microsoft SQL Server .................................................................................................................... 12  
    Microsoft SQL Server 2012 ....................................................................................................... 12  
    Microsoft SQL Server 2014 ....................................................................................................... 12  
    AlwaysOn Failover Cluster Instances ...................................................................................... 13  

**Solution architecture** ............................................................................................................... 15  
  Overview ...................................................................................................................................... 15  
  Architecture diagram ................................................................................................................... 15  
  Hardware resources ...................................................................................................................... 16  
  Software resources ....................................................................................................................... 17  

**Storage layer: EMC XtremIO and EMC VPLEX Metro** ............................................................... 18  
  Overview ...................................................................................................................................... 18  
  XtremIO storage design ............................................................................................................... 18  
    XtremIO database storage design considerations ..................................................................... 18  
    XtremIO storage design details ................................................................................................. 18  
  VPLEX Metro design ................................................................................................................... 19  
  VPLEX configuration considerations ......................................................................................... 19  
    VPLEX Witness ....................................................................................................................... 20  
    Logging volume ...................................................................................................................... 21  
    Transfer size ............................................................................................................................ 21
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPLEX consistency group</td>
<td>22</td>
</tr>
<tr>
<td>Detailed storage design for the VPLEX Metro</td>
<td>22</td>
</tr>
<tr>
<td>Microsoft SQL Server database design</td>
<td>24</td>
</tr>
<tr>
<td>Overview</td>
<td>24</td>
</tr>
<tr>
<td>SQL Server stretch cluster</td>
<td>24</td>
</tr>
<tr>
<td>Client recovery latency during failover</td>
<td>24</td>
</tr>
<tr>
<td>OLTP database profile</td>
<td>24</td>
</tr>
<tr>
<td>OLTP database design</td>
<td>25</td>
</tr>
<tr>
<td>Network layer</td>
<td>26</td>
</tr>
<tr>
<td>Overview</td>
<td>26</td>
</tr>
<tr>
<td>SAN networking best practices</td>
<td>26</td>
</tr>
<tr>
<td>IP network best practices</td>
<td>26</td>
</tr>
<tr>
<td>VMware vSphere network best practices</td>
<td>26</td>
</tr>
<tr>
<td>Physical servers and virtualization layer</td>
<td>27</td>
</tr>
<tr>
<td>Overview</td>
<td>27</td>
</tr>
<tr>
<td>Compute and storage resources</td>
<td>27</td>
</tr>
<tr>
<td>Network virtualization</td>
<td>28</td>
</tr>
<tr>
<td>Best practices for this solution</td>
<td>29</td>
</tr>
<tr>
<td>Overview</td>
<td>29</td>
</tr>
<tr>
<td>FC switch configuration</td>
<td>29</td>
</tr>
<tr>
<td>VPLEX virtualization of XtremIO storage</td>
<td>29</td>
</tr>
<tr>
<td>VPLEX sizing considerations</td>
<td>30</td>
</tr>
<tr>
<td>Business Continuity Solution Designer</td>
<td>31</td>
</tr>
<tr>
<td>Multipath consideration</td>
<td>31</td>
</tr>
<tr>
<td>Server configuration</td>
<td>31</td>
</tr>
<tr>
<td>Server configuration</td>
<td>31</td>
</tr>
<tr>
<td>ESXi server configuration</td>
<td>32</td>
</tr>
<tr>
<td>vSphere Native Multipathing configuration</td>
<td>33</td>
</tr>
<tr>
<td>Round-robin path management enablement in vCenter GUI</td>
<td>33</td>
</tr>
<tr>
<td>SQL Server failover clustering considerations in VPLEX/XtremIO environment</td>
<td>34</td>
</tr>
<tr>
<td>SQL server virtual machine design consideration for FCI</td>
<td>34</td>
</tr>
<tr>
<td>Multisite clustering</td>
<td>35</td>
</tr>
<tr>
<td>Using file share witness</td>
<td>35</td>
</tr>
<tr>
<td>Performance testing and validation</td>
<td>37</td>
</tr>
<tr>
<td>Overview</td>
<td>37</td>
</tr>
<tr>
<td>Notes on results</td>
<td>37</td>
</tr>
<tr>
<td>Test objectives</td>
<td>37</td>
</tr>
</tbody>
</table>
Test scenarios ........................................................................................................................................ 38
OLTP workload performance test .......................................................................................................... 38
  Test methodology ................................................................................................................................ 38
  Test procedure ...................................................................................................................................... 38
  Test results .......................................................................................................................................... 38
Failover test ........................................................................................................................................... 43
  Test methodology ................................................................................................................................ 44
  Test procedure ...................................................................................................................................... 44
  Test results .......................................................................................................................................... 46

Conclusion .............................................................................................................................................. 49
  Summary ............................................................................................................................................. 49
  Findings ............................................................................................................................................. 49

References .............................................................................................................................................. 50
  EMC documentation ............................................................................................................................... 50
  EMC XtremIO ......................................................................................................................................... 50
  VMware documentation ....................................................................................................................... 50
  Microsoft SQL Server documentation ................................................................................................. 50
  MSDN Library Topics .......................................................................................................................... 50
Executive summary

Business case

In today’s increasingly demanding business environments, enterprises are being driven to deliver responsive, continuously available applications that provide customers with an uninterrupted user experience. There are also higher demands on IT infrastructure performance and data availability. This environment is typically driven by:

- High-transaction workloads
- Time-critical applications and escalating service-level agreements
- Turnkey third-party applications with high sensitivity for I/O responsiveness
- Replication of application databases for use by supporting business processes such as business intelligence reporting, testing, and development
- Need for highly available architectures

In most environments, enterprises need to create copies of production data with minimal impact to the system, and safely repurpose those copies for business teams within their organization to use the data. Typically, they must wait hours or days to get access to copies of production data. This delay reduces their effectiveness for tasks such as business intelligence insight, testing and development (test/development), data integrity, validation, and auditing.

As enterprises attempt to improve data availability, issues occur when the technology solution cannot meet expectations, such as:

- Complex configuration for SQL Server environments for production, test/development, and analytics
- Limited capabilities to maintain multiple copies of databases for read and write purposes, without affecting production performance or incurring significant costs for duplicated high-performance environments
- Unwieldy backup and recovery methods with third-party tools increase costs and workloads for operations staff
- Microsoft AlwaysOn Failover Cluster Instances (FCIs) on shared storage is not available when a site is totally down

Enterprises that rely on Microsoft SQL Server must consider new approaches to meet continuing operational performance and capacity management challenges. Currently, they must consider systems that provide higher levels of performance and availability while minimizing operational costs and complexity.

Solution overview

Working together, Microsoft and EMC supply the critical components to deliver high-performance, enterprise-class availability solutions for SQL Server environments. With EMC® XtremIO™, EMC provides you with a storage solution that is optimized for extreme online transactional processing (OLTP) database performance for SQL Server and ensures that you can maximize the efficiencies of other system resources, such as CPU and memory.
The EMC VPLEX® Metro virtualized storage provides a highly available solution for a multisite cluster within Failover Cluster Instances that can also handle site failure without prolonged downtime.

Through XtremIO array-based snapshot capabilities, this solution not only provides near-instant recovery technology to minimize downtime once a database issue occurs (data loss, logical corruption, and so on), but also enables faster, easier, and more cost effective data accessibility that improves business intelligence and analytics.

XtremIO all-flash arrays with VPLEX Metro resolve database storage challenges by:

- Creating a volume with just a few clicks and enabling the entire database structure to be put into it. No planning, provisioning, or tuning steps are required.
- Automatically employing all storage system resources—solid-state drives (SSDs) and controllers—all the time.
- Scaling out the XtremIO system and increasing performance if your requirements exceed what a single XtremIO X-Brick delivers.
- Eliminating complexities by using XtremIO snapshots to manage multiple instances and copies of databases.
- Providing agility and flexibility in a test/development environment with XtremIO snapshots, which increases the productivity and quality of the applications.
- High availability in multisite clustering provided by VPLEX Metro with minimal impact on the performance of the user database.

Key results

The solution shows that the XtremIO all-flash storage array and VPLEX Metro delivers:

- Fast and simple setup with little to no storage tuning. XtremIO works as seamlessly in virtualized SQL Server environments as in physical ones, and is easy to manage and monitor.
- Support for the most demanding transactional SQL Server 2012 and SQL Server 2014 workloads in a multisite Microsoft clustered environment built on a virtualized VMware vSphere 5.5 environment.
- VPLEX virtualized XtremIO storage enables multisite clusters for SQL Server, which provides a high-availability and disaster-recovery solution with extreme performance.
- High availability enabled by VPLEX that provides near-instant recovery of production SQL Server instances during a host failure or site failure, even in TBs of data scale. This ensures the continuous service of the SQL Server databases.
Introduction

Purpose
This white paper describes a highly available and scalable solution for Microsoft SQL Server deployed in a virtualized vSphere environment with XtremIO storage virtualized by VPLEX. The white paper also demonstrates that SQL Server instances deployed on XtremIO with VPLEX are resilient during host failure or even site failure with no impact to the consolidated production server performance.

Scope
The white paper demonstrates how:

- The solution improves and enhances the performance of SQL Server 2012 and 2014 by providing new capabilities and simplifying the configuration of the environment
- XtremIO deployed with VPLEX Metro provides high availability with minimal performance impact on SQL Server multisite clustered databases

Audience
The white paper is intended for Microsoft SQL Server database administrators (DBAs), VMware administrators, storage administrators, IT architects, and technical managers responsible for designing, creating, and managing Microsoft SQL Server databases, infrastructure, and data centers.

Terminology
Table 1 lists terminology included in this white paper.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data synchronization</td>
<td>The process by which changes to a primary database are reproduced on a secondary database.</td>
</tr>
<tr>
<td>OLTP</td>
<td>Typical applications of online transaction processing (OLTP) include data entry and retrieval transaction processing.</td>
</tr>
<tr>
<td>Multisite clustering</td>
<td>A configuration where each failover cluster node is connected to a different subnet or different set of subnets. Clustering across geographically dispersed sites is sometimes referred to as stretch clusters.</td>
</tr>
<tr>
<td>Round robin</td>
<td>Round robin uses an automatic path selection policy to rotate through all available paths, enabling the distribution of load across the configured paths. Round robin can present one of the most performance-effective ways of path selection. The next available I/O path in the list is selected without any determining factor. If you had, for example, six I/Os in the queue for storage, paths 1 to 6 would be used in order.</td>
</tr>
</tbody>
</table>
Technology overview

Overview
The key technology components used in this white paper are:

- EMC XtremIO
- EMC VPLEX
- VMware vSphere
- Microsoft SQL Server

EMC XtremIO
The EMC XtremIO storage array is an all-flash system based on scale-out architecture. The system uses building blocks, called X-Bricks, which can be clustered together to grow performance and capacity as required. This solution uses two X-Bricks clustered together as a single logical storage system.

Writable snapshots
XtremIO snapshots are equivalent to production volumes in terms of performance, property, and functions, which means that a snapshot in XtremIO can be considered the same as the production volume.

Figure 1 shows how XtremIO works in an environment with a demand for large amounts of test/development and quality assurance (QA) data from a writeable snapshot.

![Figure 1. XtremIO snapshot](image-url)
**XtremIO Management Server**

XtremIO Management Server (XMS) is a standalone dedicated Linux-based server that is used to control the XtremIO system operation. The array continues operating if it is disconnected from XMS but cannot be configured or monitored.

**EMC VPLEX**

The EMC VPLEX family of products, running the EMC GeoSynchrony® operating system, provides an extensive offering of features and functionality for cloud computing.

EMC VPLEX enables users to access a single copy of their data at different geographical locations concurrently, enabling a transparent migration of running virtual machines between data centers. This capability enables transparent load sharing between multiple sites while providing the flexibility of migrating workloads between sites in anticipation of planned events. In case of an unplanned event that causes disruption of services at one of the data centers, the failed services can be restarted at the surviving site with minimal effort.

VPLEX with the GeoSynchrony operating system (OS) is a storage area network (SAN) based federated solution that removes physical barriers within a single and multiple virtualized data centers. The VPLEX platform delivers both local and distributed federation. Local federation provides the transparent cooperation of physical storage elements within a site, while distributed federation extends the concept between two locations across distance. The distributed federation is enabled by AccessAnywhere™, a technology available with VPLEX that enables a single copy of data to be shared, accessed, and relocated over distance.

The combination of a virtualized data center with the VPLEX offering enables you to:

- Move virtualized applications across data centers
- Enable workload balancing and relocation across sites
- Aggregate data centers and deliver continuous service

**EMC VPLEX clustering architecture**

VPLEX uses a clustering architecture and enables servers at multiple data centers to have concurrent read and write access to shared block storage devices.

A VPLEX cluster, as shown in Figure 2, can scale up through the addition of more engines, and scale out by connecting multiple clusters to form an EMC VPLEX Metro configuration.
VPLEX Metro supports up to two VPLEX clusters, which can be in the same site or at two different sites within synchronous distances (less than 5 ms round-trip time).

VMware vSphere is a complete and robust virtualization platform, virtualizing business-critical applications with dynamic resource pools for unprecedented flexibility and reliability. It transforms the physical resources of a computer by virtualizing the CPU, RAM, hard disk, and network controller. This transformation creates fully functional virtual machines that run isolated and encapsulated operating systems and applications.

VMware vSphere 5.5 has the following scalability and performance enhancements, which enable a virtual machine to use more resources from the hypervisor:

- Support for 62 TB virtual machine datafile (VMDK)
- Microsoft Cluster Service (MSCS) updates: VMware introduced several additional features to support MSCS, including:
  - Microsoft Windows 2012 clusters
  - “Round-robin” path policy for shared storage

---

1 A number of changes were made in vSphere 5.5 concerning the SCSI locking mechanism used by MSCS when a failover of services occurs. To facilitate this new path policy, changes
• iSCSI protocol for shared storage
• FC over Ethernet (FCoE) protocol for shared storage for round-robin support
• End-to-end support: VMware introduced 16 GB end-to-end FC support; both the HBAs and array controllers can run at 16 GB as long as the FC switch between the initiator and target supports it
• PDL AutoRemove: Introduced with vSphere 5.5, this feature automatically removes a device from a host when the device enters a PDL state
• VMware vSphere Replication Interoperability
• VMware vSphere Replication Multi-Point-in-Time Snapshot Retention
• VMware vSphere Flash Read Cache

XtremIO offers efficient enterprise storage working with VMware vSphere 5.5 cloud infrastructure.

**Microsoft SQL Server 2012**

Microsoft SQL Server 2012 is the Microsoft database management and analysis system for eCommerce, line-of-business, and data warehousing solutions.

**AlwaysOn**

SQL Server AlwaysOn presents enhanced capabilities for both specific databases and entire instances, providing flexibility to support various high availability configurations through:

• AlwaysOn FCIs
• AlwaysOn Availability Group (AAG)

**AlwaysOn Availability Groups**

AAG is a high-availability and disaster recovery solution introduced in SQL Server 2012, which enables administrators to maximize availability for one or more user databases. SQL Server instances are configured so that a single primary database or a group of primary databases can have up to four secondary database copies residing on Windows Server Failover Cluster (WSFC) nodes.

**Readable Columnstore indexes**

The Columnstore index introduced in SQL Server 2012 provides significantly improved performance for data warehousing types of queries.

SQL Server 2012 Columnstore indexes cannot be dynamically updated.

**Microsoft SQL Server 2014**

Microsoft’s release of SQL Server 2014 has several compelling features.

**In-Memory OLTP Engine**

By moving selected tables and stored procedures into memory, SQL Server 2014 can drastically reduce I/O operations and improve performance of the OLTP applications.

have been implemented that make it irrelevant which path is used to place the SCSI reservation; any path can free the reservation.
**Enhanced Windows Server 2012 Integration**

SQL Server 2014 provides improved integration with Windows Server 2012 and Windows Server 2012:

- Scales up to 640 logical processors and 4 TB of memory in a physical environment
- Scales up to 64 virtual processors and 1 TB of memory when running on a virtual machine
- Supports Windows 2012 R2 Storage Spaces feature to create tiered storage pools that improve performance
- Takes advantage of Server Message Block (SMB) 3.0 enhancements to achieve high-performance database storage on file shares

  With the new SMB Direct feature, you can use the network interface card (NIC) remote direct memory access (RDMA) feature to provide access speeds for SMB file shares nearing the access speed for local resources.

**Resource Governor Improvement**

The SQL Server 2014 Resource Governor provides a new capability to manage application storage I/O utilization. The Resource Governor can limit the physical I/Os issued for user threads in a given resource pool, enabling more predictable application performance. This can be used to limit the number of I/Os issued at the SQL Server instance boundary.

**Buffer Pool Extension**

The Buffer Pool Extension (BPE) provides a seamless integration of SSDs as a high-speed, nonvolatile random access memory (NVRAM) extension to the Database Engine’s standard buffer pool to significantly improve I/O throughput. The new buffer pool extensions can provide the best performance gains for read-heavy OLTP workloads.

**Enhancements to AlwaysOn Availability Groups**

The SQL Server 2014 AAG has been enhanced with support for additional secondary replicas and Windows Azure integration.

Readable secondary replicas in SQL Server 2014 are available for read-only workloads, even when the primary replica is unavailable.

**Updateable Columnstore indexes**

Columnstore indexes in SQL Server 2014 are updateable.

**AlwaysOn Failover Cluster Instances**

The SQL Server multisubnet failover clusters enable geographically dispersed sites to replicate SQL Server data. Clustering across geographically dispersed sites is also referred to as “stretch clusters”. Because the sites do not share storage, data replication through the storage layer enables more than one copy of available data. Therefore, the SQL Server AlwaysOn FCIs provide both a high availability and a disaster recovery solution.
Figure 3 shows a two-node, two-subnet FCI in SQL Server. The VPLEX virtualized XtremIO storage device provides the data replication mechanism between the sites in this solution.

Figure 3. Stretch cluster for Microsoft SQL Server
Solution architecture

Overview

This solution provides cost-effective performance for Microsoft SQL Server mission-critical application environments. The SQL Server 2012 and 2014 instances are deployed for virtualized production databases on an XtremIO storage array consisting of two X-Bricks. The storage for these production databases is replicated through VPLEX for high availability. In addition, virtualized test/development SQL Server instances can access directly mounted XtremIO snapshots of the production database. The snapshots can be taken from both sites.

Figure 4 shows the logical architecture of this solution.

Figure 4. Solution architecture

The architecture is composed of the following:

- **Storage layer**: Each site comprises two X-Bricks with four active storage controllers in a single XtremIO cluster (XtremIO version 2.4) with 14.94 TB of usable physical capacity, and a VPLEX Metro with four director share groups on each site. A VPLEX Witness server is configured to provide seamless zero or near-zero failover.
• **SQL Servers database layer:** Comprised of both SQL Server 2012 and SQL Server 2014 FCI cluster as production servers. SQL Server 2012 has six databases and a total of about 7 TB data. SQL Server 2014 has three databases and total of about 4 TB data. Snapshots are created on both production and remote sites, and can be mounted onto any of the mount hosts at any time, when needed. A file-share witness is configured for both SQL Server 2012 and SQL Server 2014 to monitor the cluster health from database level.

• **Network layer:** Comprised of two IP switches and two director-class SAN switches, which we² configured to produce 108 GB/s active bandwidth. The SAN switches are designed for deployment in storage networks supporting virtualized data centers and enterprise clouds. Latency is not introduced between the wide-area network (WAN) links of the two sites.

• **Physical servers and virtualization layer:** Comprised of three servers that use a total of 120 Intel E7 processor cores with 2.9 GHz processors and a total of 2 TB RAM. The rack server enables a high-performing, consolidated, virtualized approach to a Microsoft SQL Server infrastructure, resulting in deployment flexibility without the need for application modification.

The servers are installed with vSphere 5.5 and configured as a VMware ESXi cluster. The cluster is comprised of two enterprise-class production SQL Server virtual machines (SQL Server 2012 and SQL Server 2014). There are also three more standalone SQL Servers—two SQL Server 2012 and one SQL Server 2014 virtual machines. Each virtual machine is configured with 16 vCPUs and 32 GB RAM.

Table 2 lists the hardware resources used in the solution.

### Table 2. Hardware resources

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Quantity</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage array</td>
<td>2</td>
<td>EMC XtremIO consisting of two X-Bricks</td>
</tr>
<tr>
<td>Storage virtualization</td>
<td>2</td>
<td>Two EMC VPLEX Metro systems with four active storage directors</td>
</tr>
<tr>
<td>Servers</td>
<td>6</td>
<td>40 cores, 2.9 GHz processors, 512 GB RAM, including:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Two 1 Gb Ethernet (GbE) NICs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Two 10 GbE NICs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Two 8 GB FC dual-port HBAs</td>
</tr>
<tr>
<td>LAN switches</td>
<td>2</td>
<td>10 GbE, 32-port non-blocking</td>
</tr>
<tr>
<td>SAN switches</td>
<td>2</td>
<td>Two FC director-class switches with six blades</td>
</tr>
</tbody>
</table>

² In this white paper, “we” refers to the EMC Solutions engineering team that validated the solution.
Software resources  Table 3 lists the software resources used in this solution.

<table>
<thead>
<tr>
<th>Software</th>
<th>Version</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMC XtremIO</td>
<td>2.4</td>
<td>All-flash storage</td>
</tr>
<tr>
<td>EMC GeoSynchrony</td>
<td>5.2</td>
<td>Operating environment for VPLEX</td>
</tr>
<tr>
<td>VMware vSphere</td>
<td>5.5</td>
<td>Hypervisor hosting all virtual machines</td>
</tr>
<tr>
<td>VMware vCenter</td>
<td>5.5</td>
<td>Management of vSphere</td>
</tr>
<tr>
<td>Microsoft Windows</td>
<td>2012 R2</td>
<td>Operating system for database servers</td>
</tr>
<tr>
<td>Microsoft SQL Server</td>
<td>2012 SP1 Enterprise Edition</td>
<td>Database</td>
</tr>
<tr>
<td>Microsoft SQL Server</td>
<td>RTM Enterprise Edition</td>
<td>Database</td>
</tr>
<tr>
<td>Microsoft BenchCraft TPC-E Toolkit</td>
<td>1.12.0-1026</td>
<td>TPC-E-like OLTP benchmark workload tool</td>
</tr>
</tbody>
</table>
Storage layer: EMC XtremIO and EMC VPLEX Metro

Overview

This solution uses VPLEX Metro to stretch XtremIO storage over two separate sites for continuous storage availability.

All previous stated design best practices with XtremIO and VPLEX Metro apply to this environment. In addition, this section describes a few specific design considerations for VPLEX Metro working with XtremIO in a SQL Server environment.

XtremIO storage design

XtremIO uses its multi-controller scale-out design and RDMA fabric to maintain all metadata in memory.

This makes XtremIO arrays impervious to changes in workload—it does not matter what LUN sizes are used, whether there are random or sequential access patterns, or if there is locality of reference or not. The performance is always consistent and predictable.

The need for a careful, painstaking storage design for optimized performance is no longer necessary. For example, disruptive tempdb database workloads can coexist in the same LUN with its write-intensive transaction logs and still provide excellent performance. With built-in thin provisioning, storage is only allocated when it is needed. This enables you to create larger LUNs to accommodate future or unexpected growth for databases, without wasting any physical space on storage.

Most importantly, heavy metadata operations such as inline data reduction, thin provisioning allocations, and internal array copy operations are conducted entirely in memory, instantly, without impacting I/O.

XtremIO database storage design considerations

Performance is the number one consideration for tier-1 database storage design, but inherent in traditional performance storage designs are the complexities and high costs.

Database storage design typically requires free space at all levels of the storage stack, from actual data in databases, to space allocated to datafiles and log files.

With XtremIO, using thin provisioning (allocate-on-demand) and deduplication means that a 1 TB database requires less than 1 TB of allocated physical space. Operational complexities can be eliminated by allocating as much LUN space, virtual file system space, and NTFS volume space as required because storage is only allocated on demand.

XtremIO storage design details

For this solution, XtremIO is deployed in a two X-Brick cluster configured by default with XtremIO XDP to provide a physical capacity of 14.94 TB. All datafiles and log files are configured as VPLEX distributed virtual volumes and replicated to the secondary site. The two sites have similar deduplication and thin provisioning savings, as shown in Table 4. The production site has some management servers, such as vCenter Server; therefore, it uses more physical storage and has a slightly lower overall efficiency.
**Table 4. Deduplication ratio and thin provisioning savings on the sites**

<table>
<thead>
<tr>
<th>Site</th>
<th>Overall efficiency</th>
<th>Deduplication ratio</th>
<th>Thin provisioning savings</th>
<th>Volume capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production site</td>
<td>3.9:1</td>
<td>1.4:1</td>
<td>64%</td>
<td>38.8 TB</td>
</tr>
<tr>
<td>Remote site</td>
<td>5.2:1</td>
<td>1.2:1</td>
<td>76%</td>
<td>30.6 TB</td>
</tr>
<tr>
<td>Production site with five snapshots</td>
<td>15:1</td>
<td>1.8:1</td>
<td>88%</td>
<td>188.0 TB</td>
</tr>
<tr>
<td>Remote site after creating five snapshots</td>
<td>20:1</td>
<td>1.8:1</td>
<td>91%</td>
<td>181.0 TB</td>
</tr>
</tbody>
</table>

**Note:** Because of the way databases store data, low deduplication rates for single copies of databases are expected.

When we create five snapshots for all databases on the VPLEX distributed disks over the XtremIO storage, the deduplication ratio on the primary site increases. The secondary site does not change because the snapshots are not replicated to the remote site. The secondary site changes when we also create snapshots on that site.

**VPLEX Metro design**

This solution is designed for customers who require concurrent access and data mobility across two locations separated by synchronous distances.

VPLEX is a storage virtualization appliance that sits between the front-end hosts and the back-end storage array, in this case, the XtremIO X-Bricks.

VPLEX has a built-in global cache, covering the entire back-end storage farm. VPLEX offloads a lot of I/O operations from the back end. Latencies can be improved from both the direct effect of caching and the indirect effect of a more lightly loaded back end.

Although this solution tested VPLEX Metro without WAN latency, a similar architecture can be used in a one-site EMC VPLEX Local solution. Similar or better performance results can be expected with VPLEX Local providing high availability during a system upgrade or software update, during which the production host is temporarily offline, or during a host failure.

**VPLEX configuration considerations**

In order to create a virtualized device that spans two different sites, the volume used for the same virtual storage needs to be created with the same size on the XtremIO for both sites. These volumes are zoned to the VPLEX on each site. VPLEX claims these volumes, including the system database, tempdb, database and log LUNs, and a datastore LUN for raw device mappings (RDMs), VPLEX then uses the volumes to form virtualized storage for SQL Server to build the stretch cluster across the different sites.

The volumes for the SQL Server cluster configuration are required to be RDMs in a VMware environment. The datastore LUNs created as RDMs ensure that when the
cluster fails over to any site, the mappings are still valid, and the virtual machines immediately recognize the volumes.

As shown in Figure 5, the distributed device_SQL2014-DB1_1 was mapped to the device SQL2014-DB1 on the production XtremIO in VPLEX cluster-1 and to the device SQL2014-DB1-rm on the remote XtremIO in VPLEX cluster-2. On both sites, SQL Server views this storage as the virtual device device_SQL2014-DB1-1, yet they physically reside on two different XtremIO and VPLEX systems on each site.

![Figure 5. VPLEX distributed volume map](image)

**VPLEX Witness**

VPLEX Witness helps VPLEX clusters distinguish between VPLEX cluster failures and intercluster partition failures. VPLEX Witness observes health-check heartbeats to both clusters over the IP network and notifies clusters about its observations.

All distributed virtual volumes are still configured with the preferred detach rules, but, for cluster or connectivity failures, VPLEX Witness forces the majority rule to take precedence over the preference rule. This means that in order for a given VPLEX cluster to continue processing I/O operations, the cluster must either be connected to a peer VPLEX cluster or to VPLEX Witness. The static preference plays a role only in the case of an intercluster network partition when both clusters remain connected to VPLEX Witness.

VPLEX Witness is deployed as a virtual machine hosted by a VMware ESXi server.
**Note:** If both VPLEX clusters are active, or both VPLEX clusters are passive at the time of a link failure, I/O operation is suspended on both clusters to ensure data integrity. This only occurs when the preference is set to **Active Cluster Wins**.

### Logging volume

Logging volume are required for creating distributed devices in VPLEX Metro. The logging volume keeps track of blocks written during a loss of connectivity between VPLEX clusters. A 10 GB logging volume can support 320 TB of distributed storage. It only receives I/O during link outages. Because it keeps the data when two clusters are not synchronized, the safety and performance of logging volume is important. EMC recommends hosting the logging volume on XtremIO and creating a mirror on the remote site.

The best practices for configuring logging volumes on a VPLEX are:

- Create one logging volume for each cluster.
- The data protection capabilities provided by the XtremIO storage array ensure the integrity of the logging volumes.
- Create a mirror on a different array for further redundancy.
- Configure at least 10 GB of logging volume space. Because the physical storage on XtremIO is only allocated when needed, there is no waste, even if you create a larger volume space than is actually needed.

The VPLEX rebuild process distributes volume by synchronizing it on the two clusters. Distributed volumes on two clusters will become identical again after a disaster through rebuild.

### Transfer size

Transfer size is the size of the region in the cache used to service the migration. The area is globally locked, read at the source, and written at the target. Transfer size can be as small 40k and as large as 128 MB, and must be a multiple of 4k. Consider the following for transfer sizing:

- A larger transfer size results in higher performance for the migration with impact on the front-end I/O operations. This is especially true for VPLEX Metro migrations.
- A smaller transfer size results in lower performance for the migration, but creates less impact on front-end I/O operations and response times for hosts.
- Set a large transfer size for migrations when data protection or migration performance is important. Set a smaller transfer size for migrations when front-end storage response time is more important.

We used the default 128k for this solution to ensure the best performance after the failover.
**VPLEX consistency group**

VPLEX consistency groups aggregate volumes to enable the application of a common set of properties to the entire group. All consistency groups guarantee a crash-consistent image of their member virtual volumes.

There are two types of consistency group:

- **Synchronous consistency group** aggregates local and distributed volumes on VPLEX Local and on VPLEX Metro systems that are separated by 5 ms or less of latency.

- **Asynchronous consistency group** aggregates distributed volumes on VPLEX Geo systems separated by 50 ms or less of latency.

This solution uses synchronous consistency groups for the VPLEX Metro setup.

For more information, refer to the *EMC VPLEX Metro Witness Technology and High Availability TechBook*.

**Detailed storage design for the VPLEX Metro**

In this solution, database data and log file LUNs are configured as RDM and created as distributed volumes. Virtual machine OS volumes are created as VPLEX local volumes and do not need to be replicated to the remote site. A small amount of storage is required for VPLEX management and logging, which is described in Table 5.

### Table 5. Storage design for Microsoft SQL Server cluster with VPLEX distributed storage on XtremIO

<table>
<thead>
<tr>
<th>Volume name</th>
<th>Volume purpose</th>
<th>Volume type</th>
<th>LUN size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL_OS</td>
<td>Microsoft Windows 2012 R2 OS and SQL Server software installation volume, which is used for multiple virtual machines such as VMDK on the same datastore</td>
<td>VPLEX local</td>
<td>1 TB</td>
</tr>
<tr>
<td>SQL_DB</td>
<td>Microsoft SQL Server database datafile volumes</td>
<td>VPLEX distributed</td>
<td>2 TB</td>
</tr>
<tr>
<td>SQL_log</td>
<td>Microsoft SQL Server database log file volumes</td>
<td>VPLEX distributed</td>
<td>500 GB</td>
</tr>
<tr>
<td>System/tempdb</td>
<td>Microsoft SQL Server system/tempdb volumes</td>
<td>VPLEX distributed</td>
<td>1 TB</td>
</tr>
<tr>
<td>VPLEX metadata and backups</td>
<td>Meta data for VPLEX Backup volumes for VPLEX metadata</td>
<td>VPLEX meta</td>
<td>80 GB</td>
</tr>
<tr>
<td>VPLEX logging</td>
<td>Logging volume for VPLEX metro</td>
<td>VPLEX logging</td>
<td>10 GB</td>
</tr>
<tr>
<td>VPLEX ESXi common datastore</td>
<td>Common datastore accessible from all servers in the cluster to ensure the datastore mapping on ESXi server when any VPLEX volume is failed over to that server.</td>
<td>VPLEX distributed</td>
<td>120 GB</td>
</tr>
</tbody>
</table>
For the production databases, volumes are created and presented to the virtual machine for use with the Microsoft SQL Server cluster virtual machines, as shown in Table 6.

Table 6. **Volume/LUN assignment for OLTP Database**

<table>
<thead>
<tr>
<th>Volume</th>
<th>Volume size</th>
<th>Volume type</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>120 GB</td>
<td>VMDK on OS LUN/VMFS volume</td>
</tr>
<tr>
<td>SQL Server</td>
<td>2 TB</td>
<td>RDM</td>
</tr>
<tr>
<td>SQL Server log</td>
<td>500 GB</td>
<td>RDM</td>
</tr>
<tr>
<td>SQL Server installation and systems databases/tempdb</td>
<td>1 TB</td>
<td>RDM</td>
</tr>
</tbody>
</table>

**Note:** Windows Failover Clustering—used in this solution—requires RDMs when running in virtual-machine clustering (to support SCSI-3 reservations).
Microsoft SQL Server database design

Overview
In this solution, two virtualized instances with transactional OLTP databases (one on the Microsoft SQL Server 2012 cluster and one on the Microsoft SQL Server 2014 cluster) were created on a vSphere high availability (HA) cluster.

SQL Server stretch cluster
We configured both SQL Server 2012 and SQL Server 2014 as a multisubnet failover cluster (stretch cluster) over the two different sites.

Client recovery latency during failover
A multisubnet FCI by default enables the `RegisterAllProviders/IP` cluster resource for its network name. Both the online and offline IP addresses of the network name are registered at the DNS server. The client application retrieves all registered IP addresses from the DNS server and attempts to connect to the addresses either in order or in parallel. This means that client recovery time in multisubnet failovers no longer depend on DNS update latencies.

By default, the client tries the IP addresses in order. When the client uses the optional `MultiSubnetFailover=True` parameter in its connection string, it tries the IP addresses simultaneously and connects to the first server that responds. This helps minimize the client recovery latency when failovers occur.

For more information, refer to the MSDN Library Topics *AlwaysOn Client Connectivity (SQL Server)* and *Create or Configure an Availability Group Listener (SQL Server)*.

Note: To ensure that client applications work optimally with multisubnet FCI in SQL Server 2014, adjust the connection timeout in the client connection string by 21 seconds for each additional IP address. This ensures that the reconnection attempt does not time out before it cycles through all the IP addresses in the FCI.

OLTP database profile
Table 7 lists the SQL Server database profile for the solution.

Table 7. Database profile for OLTP database

<table>
<thead>
<tr>
<th>Property</th>
<th>SQL Server 2012</th>
<th>SQL Server 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database type</td>
<td>OLTP (transactional)</td>
<td>OLTP (transactional)</td>
</tr>
<tr>
<td>Database size</td>
<td>Total: 5 TB</td>
<td>Total: 2.25 TB</td>
</tr>
<tr>
<td>Microsoft SQL Server databases</td>
<td>1 x 2 TB, 1 x 1 TB, 1 x 750 GB, 2 x 500 GB, 1 x 250 GB</td>
<td>1 x 1 TB, 1 x 750 GB, 1 x 500 GB</td>
</tr>
<tr>
<td>Memory for SQL Server</td>
<td>32 GB</td>
<td>32 GB</td>
</tr>
<tr>
<td>Workload profile</td>
<td>OLTP workload simulated by Microsoft BenchCraft</td>
<td>OLTP workload simulated by Microsoft BenchCraft</td>
</tr>
<tr>
<td></td>
<td>Read/write ratio: 90/10</td>
<td>Read/write ratio: 90/10</td>
</tr>
<tr>
<td>Average data block size</td>
<td>8 KB</td>
<td>8 KB</td>
</tr>
</tbody>
</table>
With XtremIO virtual provisioning, the database storage design can be simplified. In this solution, we created uniform volumes for the database and log file LUNs for easy management without wasting physical storage. The storage is allocated as needed.

As Table 6 on page 23 shows, we used six 2 TB database volumes to store the relevant database data files, six 500 GB volumes for transaction log files, and a 1 TB volume for system database tempdb files for the SQL Server 2012 databases. We used three 2 TB database volumes, three 500 GB log volumes, and one 1 TB volume to store the relevant files for the SQL Server 2014 databases.

Table 8 and Table 9 list the OLTP database actual LUN design for the solution.

**Table 8. OLTP database actual LUN design detail for SQL Server 2012**

<table>
<thead>
<tr>
<th>Detail</th>
<th>Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database name</td>
<td>DB_01</td>
</tr>
<tr>
<td>Actual database size</td>
<td>750 GB</td>
</tr>
<tr>
<td>LUN size</td>
<td>2 TB</td>
</tr>
<tr>
<td>Actual log size</td>
<td>350 GB</td>
</tr>
<tr>
<td>Log LUN size</td>
<td>500 GB</td>
</tr>
<tr>
<td>Total data and log size</td>
<td>7.2 TB</td>
</tr>
<tr>
<td>Total LUN size</td>
<td>16 TB</td>
</tr>
</tbody>
</table>

**Table 9. OLTP database actual LUN design detail for SQL Server 2014**

<table>
<thead>
<tr>
<th>Detail</th>
<th>Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database name</td>
<td>DB_01</td>
</tr>
<tr>
<td>Actual database size</td>
<td>750 GB</td>
</tr>
<tr>
<td>LUN size</td>
<td>2 TB</td>
</tr>
<tr>
<td>Actual log size</td>
<td>350 GB</td>
</tr>
<tr>
<td>Log LUN size</td>
<td>500 GB</td>
</tr>
<tr>
<td>Total data and log size</td>
<td>3.7 TB</td>
</tr>
<tr>
<td>Total LUN size</td>
<td>8.5 TB</td>
</tr>
</tbody>
</table>

**Note:** This design is based on our test workload. In a production environment, database size, especially log file and tempdb sizes, can vary, depending on the type of transactions and queries that are running on those databases.
## Network layer

**Overview**

This section describes the network details used in this solution for SAN and IP network configuration, and for ESXi Server networks. When deploying a virtualized database solution, such as Microsoft SQL Server, EMC recommends that you ensure both compute and network redundancy at all levels when designing networking fault tolerance.

**SAN networking best practices**

EMC recommends you use the following SAN network best practices:

- Use at least 8 Gb/s FC switches and HBA ports.
- Use multiple HBAs on the ESXi servers and at least two SAN switches to provide multiple redundant paths between the server and the VPLEX.
- Use multiple paths from the XtremIO cluster front end to the VPLEX back end.
- Zone each FC port from the database servers to all ports on the XtremIO X-Bricks for high availability and performance.

**IP network best practices**

EMC recommends that you use the following IP network best practices:

- Use multiple network cards and switches for network redundancy.
- Use 10 GbE for network connection, if available.
- Use virtual local area networks (VLANs) to logically group devices that are on different network segments or sub-networks.

**VMware vSphere network best practices**

Networking in virtual environments requires more considerations for traffic segmentation, availability, and throughput in addition to the best practices followed in a physical environment.

This solution was designed to efficiently manage multiple networks and redundancy of network adapters on ESXi hosts. The key best practice guidelines are to:

- Separate infrastructure traffic from virtual machine traffic for security and isolation.
- Use the VMXNET3 family of paravirtualized network adapters.
- Aggregate physical network cards for network redundancy and performance, for example, use pairs of physical NICs per server/vSwitch, and uplink each physical NIC to separate physical switches.

For more information on networking with vSphere, refer to the instructions in *VMware vSphere Networking*. 

---

EMC Availability for Extreme Performance of Microsoft SQL Server
EMC XtremIO, EMC VPLEX, VMware vSphere
White Paper

---

EMC 2
Physical servers and virtualization layer

Overview

The choice of a server platform for a virtualized infrastructure is based on both the supportability of the platform and the technical requirements of the environment. In production environments, it is essential that the servers used have:

- Sufficient processors and memory to support the required number and workload of the virtual machines.
- Sufficient connectivity, both Ethernet and FC, to enable redundant connectivity to the IP and storage network switches.
- Sufficient capacity to withstand a server failure and support failover of the virtual machines.

In this test environment, three physical servers running vSphere ESXi 5.5 are configured as a vSphere HA cluster. Five virtual machines are created on this vSphere cluster, two of which are configured to create virtualized Microsoft SQL Server database virtual machines. The other three virtual machines are created as Test/Development instances that can be used to mount various snapshots for repurposing (Test/Dev).

Compute and storage resources

EMC recommends that you implement the following VMware compute resource best practices as explained in the Microsoft SQL Server Databases on VMware Best Practices Guide:

- Use Non-Uniform Memory Access (NUMA) on the ESXi servers, a computer architecture in which memory located closer to a particular processor is accessed with less delay than memory located farther from that processor.
- Allocate virtual machine memory (vRAM) in a virtual machine to be less than or equal to the local memory accessed by the NUMA node (processor).
- Install VMware Tools, including several utilities that enhance the performance of the virtual machine’s guest operating system and improve the ability to manage the virtual machine.
- Configure the virtual machine memory reservations to be, at a minimum, the size of the Microsoft SQL Server and operating system overhead.
- Microsoft SQL Server only supports RDM for clustering, so use RDM in ESXi virtual machine for database and log files that needs to failover in an MSCS clustering.
- Configure multiple paravirtualized SCSI (PVSCSI) controllers for the database volumes. Using multiple virtual SCSI controllers enables the execution of several parallel I/O operations inside the guest operating system.
Network virtualization

On each ESXi server, we created two standard vSwitches with a common configuration as listed in Table 10.

**Table 10. vSwitch configuration**

<table>
<thead>
<tr>
<th>Name</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>vSwitch0</td>
<td>Management and public virtual machine traffic</td>
</tr>
<tr>
<td>vSwitch1</td>
<td>Fault tolerant configuration for Microsoft SQL Server Cluster interconnect</td>
</tr>
</tbody>
</table>

Each virtual machine was assigned two vNICs (1 GbE and 10 GbE) using the high performance VMXNET3 driver. The 1 GbE vNIC was mapped to vSwitch0 to deliver public traffic. The 10 GbE vNIC was mapped to vSwitch1 to deliver Microsoft SQL Server interconnect traffic.
Best practices for this solution

Overview
XtremIO enables extremely high I/O loads on a single storage system. With the XtremIO balanced architecture combined with performance, inline data reduction, and virtually provisioned storage, many of the fine-tuning and configuration practices that a traditional storage array requires are no longer needed.

To take full advantage of the high throughput that XtremIO storage provides, the entire connectivity stack must be configured for extreme performance, from optimizing queue depths on hosts to the number of available FC paths that must be considered, so that enough I/O can be pushed through VPLEX towards the XtremIO system.

FC switch configuration
For an XtremIO dual X-Brick cluster, a host can have up to eight paths per device. Figure 6 shows the logical connection schemes for eight paths that are configured with a dual-engine VPLEX Metro back end and connected to the host through the VPLEX front end.

![Figure 6. XtremIO dual X-Brick and VPLEX FC switch configuration for the ESXi host](image)

VPLEX virtualization of XtremIO storage
After you connect the VPLEX with XtremIO on the SAN, two 80 GB meta volumes and two meta volume backups need to be created (total of four 80 GB volumes) on the XtremIO cluster on each site. For the VPLEX Metro set up, a 10 GB logging volume needs to be created and presented to the VPLEX on that site, as shown in Figure 7.

![Figure 7. XtremIO dual X-Brick FC switch configuration](image)

To create a valid distributed volume, the VPLEX virtual volume LUN number on both clusters must be the same for the devices that forms the distributed volume, as shown in Figure 8.
Storage sizing is important when you use VPLEX Metro to provide data virtualization for XtremIO storage.

Figure 9 shows the maximum IOPS that a single engine/two-director, dual-engine/two-director, or dual-engine/four-director VPLEX can support for 4 KB I/O operations. In our testing in the SQL Server environment, which consisted mainly of 8 KB I/O operations, the maximum throughput for the dual-engine/four-director VPLEX Metro configuration closely matched the 4 KB throughput (same total I/O throughput for the MB/s).

Table 11 shows sample workloads for the peak and average front-end IOPS for 8 KB I/O size on different models of VPLEX. (Performance might vary depending on the WAN latency and SAN configuration.)
Table 11. Sample workloads

<table>
<thead>
<tr>
<th>VPLEX front-end IOPs</th>
<th>Peak workload</th>
<th>Average workload</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Write only</td>
<td>Read only</td>
</tr>
<tr>
<td>Single engine</td>
<td>35k</td>
<td>50k</td>
</tr>
<tr>
<td>Double engine</td>
<td>55k</td>
<td>100k</td>
</tr>
<tr>
<td>Four engine</td>
<td>100k</td>
<td>240k</td>
</tr>
</tbody>
</table>

For a SQL Server with mostly 8 KB random I/O operations and a 90:10 read/write ratio, one dual-engine VPLEX Metro setup can support a peak workload of about 90,000 IOPS with reasonable performance for XtremIO storage.

The actual size of the average front-end IOPs that a VPLEX system can sustain depends on specific environment, configuration, workload, and latency requirements of the application. For predictable, consistent application performance, EMC recommends that you do not constantly reach the maximum of the VPLEX.

**Business Continuity Solution Designer**

The Business Continuity Solution Designer (BCSD) is an independent tool that can be used to estimate the size of VPLEX system requirements based on the application workload.

**Note:** To perform a sizing exercise with the BCSD, engage your EMC account team.

**Multipath consideration**

Managing fewer paths for a LUN makes VPLEX more efficient, which means better performance. Having load balance from multiple paths provides better throughput for the XtremIO X-Bricks. You must consider and balance these two requirements.

With two directors instead of four for a VPLEX, you have performance gains of about 10 to 20 percent for a specific LUN. In environments that need extremely high performance, the VPLEX can be configured to use two directors instead of four directors with reduced high availability for performance gain.

In this solution, we opted to use all four directors for optimum high availability. The performance is still significantly better when compared with any traditional storage array.

**Server configuration**

To optimize performance to extreme levels, hosts accessing the XtremIO storage array must be configured to enable higher I/O throughout instead of using the default settings.

**Server configuration**

Most server default HBA throttle settings are not optimized for the high throughput that a flash array provides. Therefore, it is important to choose the highest value for this setting for the server so it will not limit the I/O throttle.
To adjust HBA I/O throttle of the Cisco UCS HBA in this solution:

1. In UCSM navigation, under Server, select Inventory.
2. Select Cisco VIC Adapters.
3. Navigate to vHBA Properties.
4. Set I/O Throttle Count to 1024, as shown in Figure 10.

5. For other server types, adjust similar settings accordingly.

**ESXi server configuration**

To optimally configure the ESXi host for XtremIO storage (for vSphere 5.5):

1. On vSphere, adjust the **HBA queue depth** through the ESX command-line interface (CLI). The queue depth setting controls the amount of outstanding I/O requests per a single path.

For optimal operation with XtremIO storage, follow the HBA vendor and server vendor recommendations. As a general rule, you should set the queue depth to the highest allowed by the HBA manufacturer (for example, 256).

**Note:** For more information about adjusting HBA queue depth with ESX, refer to VMware KB article 1267.
2. Set the **SchedQuantum** (to 64) and **DiskMaxIOSize** (to 4096) parameters:
   ```
esxcfg-advcfg -s 64 /Disk/SchedQuantum
esxcfg-advcfg -s 4096 /Disk/DiskMaxIOSize
   ```

3. Obtain the NAA for the XtremIO LUNs presented to the ESX host and locate the NAA of the distributed VPLEX encapsulated XtremIO volume (usually displayed as **EMC Fibre Channel Disk naa.6000144000000103039d8f1bd55a98**):
   ```
esxcli storage nmp path list
   ```

4. Run the following command to set **SchedNumReqOutstanding** for the device to its maximum value (256):
   ```
esxcli storage core device set -d naa.xxx -O 256
   ```

**vSphere Native Multipathing configuration**

VPLEX supports the VMware vSphere Native Multipathing (NMP) technology. For best performance, EMC recommends that you configure the native vSphere multipathing for VPLEX encapsulated XtremIO volumes using these steps:

1. Set the native round-robin path selection policy on VPLEX distributed volumes presented to the ESXi host.
2. Set the vSphere NMP round-robin path switching frequency to VPLEX distributed volumes from the default value (1,000 I/O packets) to 1.

These settings ensure optimal distribution and availability of load between I/O paths to XtremIO storage.

---

**Note:** Use the ESX command line to adjust the path switching frequency of vSphere NMP round robin.

---

To set the vSphere NMP round-robin configuration, use one of these options:

- Per volume, using vSphere Client, for each host where the volume is presented
- Per volume, using ESX command line, for each host where the volume is presented
- Per host for all XtremIO volumes, presented to the host using ESX command line

If EMC PowerPath®/VE is used for ESXi, PowerPath/VE treats VPLEX distributed devices as generic. Enabling generic loadable array module (LAM) support enables PowerPath/VE to recognize and manage VPLEX distributed devices.

**Round-robin path management enablement in vCenter GUI**

On each virtual machine, the database storage LUNs were added from the VPLEX as RDM and spread across four PVSCSI controllers to balance I/O. The LUNs for OS and SQL Server software installations are configured as VMDK so that the low I/O storage LUNs can share the same volume on VPLEX.
The I/O intensive database LUNs need to be configured as **Round Robin (VMware)** in path management, as shown in Figure 11, if they are not managed by PowerPath.

**Figure 11. Storage device path management configuration**

SQL Server failover clustering considerations in VPLEX/XtremIO environment

A Windows failover cluster consists of a group of independent computers that work together to increase the availability and scalability of clustered roles. These clustered roles contain applications and services that are designed to be highly available.

The clustered servers (called nodes) are physically connected by cables and logically grouped by failover cluster software. If one or more of the cluster nodes fails, the application or service moves (or fails over) to alternate nodes to continue providing those services without disruption. This is done by proactively monitoring each cluster resource to verify that it is working properly. If there is an issue, those targeted resources are restarted or moved to another node.

**SQL server virtual machine design consideration for FCI**

In this solution, two SQL Servers are deployed as Failover Cluster Instances (FCIs) across physical machines (CABs). This protects against software or hardware failure on the physical machine by placing the cluster nodes on separate ESXi hosts. This requires configuring the virtual machine-virtual machine anti-affinity rule for SQL Server virtual machines.

To ensure that the anti-affinity rules are strictly applied, set an advanced option for vSphere DRS. Setting the advanced option **ForceAffinePoweron** to 1 enables strict enforcement of the anti-affinity rules.

EMC also recommends that you apply vSphere DRS groups and virtual machine-host anti-affinity rules to all the virtual machines across both VPLEX clusters.
Table 12 lists some of the settings and configurations that you need to set for the virtual machine.

Table 12. SQL Server virtual machine settings for clustering

<table>
<thead>
<tr>
<th>Settings</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HKEY_LOCAL_MACHINE\System\CurrentControlSet\Services\Disk\TimeOutValue 60 seconds or more</td>
<td>I/O timeout value needs to be set to a higher number for clustering in all the virtual machine cluster nodes.</td>
</tr>
<tr>
<td>esxcli storage core device setconfig -d naa.id --perennially-reserved=true</td>
<td>Set all virtual devices to be clustered and perennially reserved in each ESXi server to improve the rescan and ESXi reboot speed.</td>
</tr>
</tbody>
</table>

Multisite clustering
Using the VPLEX in SQL Server multisite clusters is critical for ensuring continuous operations and availability. When combining MSCS and VPLEX technologies, you enable proactive operations such as disaster avoidance, which helps prevent data corruption or inconsistent results to clients during specific disaster recovery and application or service tolerance scenarios.

VPLEX enables automated restart of application resources on cluster nodes that are connected to VPLEX virtualized storage in surviving sites. This is done through a combination of FCI policies and VPLEX Witness settings, which are described in more detail in Using file share witness.

Because of their extreme disaster tolerance, multisite clusters offer a high-availability and disaster-recovery solution. The automatic failover of SQL Server multisite clusters means that your SQL Server instances and the data being managed by FCI is available within moments of a failure of your primary site.

Using file share witness
With multisite failover clusters powered by VPLEX, the quorum refers to the number of “votes” that the failover cluster must equate to form a majority. This includes all the Windows Failover Cluster nodes and a file share witness located at a third site (possibly with VPLEX Witness) to store a copy of the cluster database across all failover nodes.

VPLEX enables multisite failover clusters to be deployed in a way that automates the failover of applications in situations where the following occurs:
- Communication has failed between sites
- Complete site failure prevents applications from running

The node and file share witness quorum configuration for Windows Failover Cluster is the preferred choice for multisite clustering with VPLEX and Microsoft³. In this configuration, a file share acts as a witness and provides a tie-breaking vote for the

³ Configuration recommended in the Microsoft TechNet Library topic Configure and Manage the Quorum in a Windows Server 2012 Failover Cluster.
votes provided by the multisite cluster nodes. With the addition of this third-site file share witness, Windows Failover Cluster totals all the votes, which means that connectivity can be lost by any of the nodes (or the witness itself), but the cluster continues functioning. As shown in Figure 12, when the first site fails, for any reason, the file share witness has enough votes to enable Site-2 for the active server.

![Node and file share witness failure scenario](image)

**Figure 12. Node and file share witness failure scenario**

**Note:** The file share witness keeps track of which node has the most current version of the cluster database, but does not keep a copy of that database. VPLEX Distributed Devices enable the use of a shared disk across data centers that enables you to use the disk witness model so that you always have a backup of the cluster database.
Performance testing and validation

Overview

The purpose of this testing is not to showcase raw performance ceiling numbers of any of the compute, switch, or storage elements of this solution. The purpose is to showcase, through scaling, how business critical, transactional workloads can continue to be easily serviced by XtremIO and VPLEX while all elements, including storage, stay within the “green zone”, that is, an area of utilization and latencies that is healthy and sustainable for production workloads.

The OLTP workloads were generated using a Microsoft Partner toolkit that creates a TPC-E-like workload. This toolkit, based on the BenchCraft TPC-E Toolkit, was used to simulate realistic OLTP workloads in this solution.

The system performance metrics (IOPS, transactions per second (TPS), and latency) were collected at the server/database and storage levels.

Notes on results

Test results are highly dependent upon workload, specific application requirements, and system design and implementation. Relative system performance will vary as a result of these and other factors. Therefore, this workload should not be used as a substitute for a specific customer application benchmark when critical capacity planning and/or product evaluation decisions are contemplated.

All performance data contained in this report was obtained in a rigorously controlled environment. Results obtained in other operating environments may vary significantly.

EMC does not warrant or represent that a user can or will achieve similar performance expressed in transactions per minute.

Note: The database metric TPS is described and used within our test results. As transactions differ greatly between database environments, these figures should only be used as a reference and for comparative purposes within these test results.

Test objectives

The overall test objectives were to demonstrate:

- The high performance achieved when virtualized Microsoft SQL Server clustered databases were run on VPLEX Metro distributed devices with XtremIO.
- How VPLEX Metro encapsulated XtremIO with a Microsoft SQL Server cluster provided an efficient high availability solution for a SQL Server database.
- The minimal impact any storage, host, or site failure had on user input because the stretched cluster started on the remote site immediately after detecting a failure.
- The performance was the same or slightly better (if both production and remote VPLEX/XtremIO were utilized) after a failover.
Test scenarios

The following scenarios were tested and are described in more detail in subsequent sections:

- OLTP workload performance test
- Failover test with simulated host or site failure

OLTP workload performance test

This test was used to measure the performance of the whole environment with both SQL Server 2012 and SQL Server 2014 database workloads.

This test also showed how an XtremIO system dealt with an unexpected system failure in a production environment and continued to show stable performance after failover to a remote site using VPLEX encapsulated XtremIO storage.

Test methodology

Microsoft BenchCraft was used to generate the OLTP workload to drive high physical random I/O from a database platform.

We ran the fixed number of concurrent users for each database with the same set of OLTP queries simultaneously against all SQL Server database in the environment, then measured the performance statistics. During the test, the number of concurrent users was controlled so that we could generate a specific level of IOPS.

We simulated host failure and site failure to measure the time it took for the environment to recover from the remote site and the performance of the workload after failover.

Test procedure

The test was run with a full workload on two production clusters, as detailed in Table 13.

Table 13. Test workload sequence for full system load

<table>
<thead>
<tr>
<th>Database name</th>
<th>Database size</th>
<th>SQL Server cluster</th>
<th>Workload (No. of users/maximum transaction rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB_01</td>
<td>750 GB</td>
<td>SQL Server 2014</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_01</td>
<td>750 GB</td>
<td>SQL Server 2012</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_02</td>
<td>500 GB</td>
<td>SQL Server 2014</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_02</td>
<td>500 GB</td>
<td>SQL Server 2012</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_03</td>
<td>1 TB</td>
<td>SQL Server 2014</td>
<td>10/200</td>
</tr>
<tr>
<td>DB_03</td>
<td>1 TB</td>
<td>SQL Server 2012</td>
<td>10/200</td>
</tr>
<tr>
<td>DB_04</td>
<td>2 TB</td>
<td>SQL Server 2012</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_05</td>
<td>250 TB</td>
<td>SQL Server 2012</td>
<td>5/200</td>
</tr>
<tr>
<td>DB_06</td>
<td>1 TB</td>
<td>SQL Server 2012</td>
<td>5/200</td>
</tr>
</tbody>
</table>
In addition, we also mounted snapshots directly to the mount host of both the SQL Server 2012 cluster and the SQL Server 2014 cluster volume on XtremIO. An additional workload simulating the QA/development and business analysis was added to the mount host to simulate a busy customer environment, as detailed in Table 14.

**Table 14. Test workload sequence for mounted snapshots workload**

<table>
<thead>
<tr>
<th>Snapshot database name</th>
<th>Database size</th>
<th>Source SQL Server cluster</th>
<th>Workload (No. of users/maximum transaction rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB_01</td>
<td>750 GB</td>
<td>SQL Server 2014</td>
<td>15/200</td>
</tr>
<tr>
<td>DB_01</td>
<td>750 GB</td>
<td>SQL Server 2012</td>
<td>15/200</td>
</tr>
<tr>
<td>DB_03</td>
<td>1 TB</td>
<td>SQL Server 2012</td>
<td>10/200</td>
</tr>
<tr>
<td>DB_04</td>
<td>2 TB</td>
<td>SQL Server 2012</td>
<td>15/200</td>
</tr>
</tbody>
</table>

**Test results**

The entire system generated over 4,200 TPS with close to 200,000 IOPS on the SQL Server side, as shown in Table 15.

**Table 15. OLTP workload performance test results**

<table>
<thead>
<tr>
<th>Performance</th>
<th>Production site</th>
<th>Remote site</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Production SQL Servers</td>
<td>Test/dev SQL Servers</td>
</tr>
<tr>
<td>CPU usage</td>
<td>23–52%</td>
<td>30–64%</td>
</tr>
<tr>
<td>Total IOPS</td>
<td>201k</td>
<td>83k</td>
</tr>
<tr>
<td>Details</td>
<td>82k</td>
<td>119k</td>
</tr>
</tbody>
</table>

We added the workload on mounted snapshots for the database to simulate development work directly from XtremIO to separated virtual machines in the virtualized environment. XtremIO was fully utilized and delivered a total of almost 190,000 IOPS. With VPLEX caching also providing about 15,000 IOPS for read, the SQL Server application side operated with about 200,000 IOPS.

Overall, the average latency remained low for the XtremIO array, while multiple SQL Server database workloads generated more I/O into the system.

When all nine databases were fully loaded, array latency remained under 1 ms for the 8 KB I/O operations in the XtremIO array. The host average disk latency ranged from under 1 ms to under 4 ms.

With the additional workload from mounted-snapshot databases, the host average disk latency was not affected. The mounted-snapshot databases that mapped directly to the ESXi servers running SQL Server virtual machines delivered on average about 1 to 2 ms latency over the heavy workload.
The entire system generated over 4,200 TPS with close to 200,000 IOPS on the SQL Server side.

**SQL Server disk I/O**

The disk I/O performance was very similar for SQL Server 2012 and SQL Server 2014.

As shown in Figure 13, SQL Server 2014 latency was about 3 to 4 ms and sustained a total of about 200,000 IOPS (total including SQL Server 2014, SQL Server 2012). The entire snapshot workload was posted on the virtual machines with XtremIO Snapshot directly mounted on the test/development virtual machines.

![Figure 13. SQL Server disk I/O performance](image)

**Log file LUN disk I/O latencies are similar in both versions of SQL Server, with much lower IOPS in SQL Server 2012 (due to fewer transactions).**

Tempdb had significantly less IOPS in SQL Server 2012, which indicates that SQL Server 2014 may have improved the transaction performance by optimizing the execution plan and therefore making more use of tempdb. The tempdb latency, however, was also much lower in SQL Server 2014, making SQL Server 2014 a better choice for performance-sensitive and I/O-intensive SQL Server databases.

**VPLEX performance**

In the fully loaded environment, we pushed the total VPLEX IOPS up to about 120,000 for the four-director share group VPLEX Metro configuration (with its primary site VPLEX CPU usage pushed up to 100 percent for stress testing).

In the testing shown in Figure 14, with the primary site VPLEX CPU usage about 90 percent, the solution maintained a sustained 90,000 IOPS from the VPLEX front end, with about 2 to 3 ms read latency and 3 to 4 ms write latency on the SQL Server side.
Figure 14.  VPLEX primary site running 90 percent CPU with about 90,000 IOPS

The write latency introduced by VPLEX was about 1 to 1.6 ms in this environment, as shown in Figure 15.

Figure 15.  VPLEX introduced about 1 ms write latency on average
**Note:** This environment did not introduce WAN latency between the VPLEX sites. The write latency was expected to increase with the increase in WAN distance latency.

For read latency, VPLEX Metro without WAN latency added about 200 µs on top of the XtremIO storage. The front end provided a total of 90,000 IOPS. About 15,000 IOPS for read came from the VPLEX cache. The back-end XtremIO provided about 75,000 IOPS. The offloaded read I/O in the VPLEX cache contributed to the low read latency increase from VPLEX.

**XtremIO system performance**

As shown in Figure 16, the VPLEX encapsulated XtremIO array can sustain a high I/O requirement with low latency with multiple concurrent SQL Server database workloads in a VPLEX environment.

![XtremIO IOPS performance and latency](image-url)
When we pushed the full system load, XtremIO provided extremely high throughputs through the VPLEX distributed volume with very low latency, as shown in Table 16.

### Table 16. Production workload

<table>
<thead>
<tr>
<th>Latency and IOPS</th>
<th>Production site</th>
<th>Remote site</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Production workload on VPLEX distributed volume</strong></td>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
</tr>
<tr>
<td><strong>Latency and IOPS</strong></td>
<td><strong>Production site</strong></td>
<td><strong>Remote site</strong></td>
</tr>
<tr>
<td>Current Total</td>
<td>Performance</td>
<td>Most Active</td>
</tr>
<tr>
<td></td>
<td>Bandwidth</td>
<td>IOPS</td>
</tr>
<tr>
<td>64KB</td>
<td>872</td>
<td>873</td>
</tr>
<tr>
<td>32KB</td>
<td>642</td>
<td>673</td>
</tr>
<tr>
<td>16KB</td>
<td>484</td>
<td>460</td>
</tr>
<tr>
<td>8KB</td>
<td>339</td>
<td>353</td>
</tr>
<tr>
<td>4KB</td>
<td>211</td>
<td>204</td>
</tr>
<tr>
<td>2KB</td>
<td>011</td>
<td>013</td>
</tr>
<tr>
<td>1KB</td>
<td>009</td>
<td>009</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>77,601 IOPS</td>
<td>2,358 IOPS</td>
</tr>
</tbody>
</table>

| Production workload on VPLEX distributed volume and QA/development workload on XtremIO volume | ![Image 3](image3.png) | ![Image 4](image4.png) |
| Current Total | Performance | Most Active |
| | Bandwidth | IOPS | Latency |
| 64KB | 1,072| 1,073 | |
| 32KB | 693| 693 | |
| 16KB | 493| 493 | |
| 8KB | 293| 293 | |
| 4KB | 193| 193 | |
| 2KB | 193| 193 | |
| 1KB | 193| 193 | |
| **Total** | 184,926 IOPS | 2,300 IOPS |

**Note:** The majority I/O performance for VPLEX replication is about 32 KB.

The overall latency on VPLEX for the majority of the I/O operations (size 8 KB) corresponding to OLTP workload was on average around 300 to 500 µs/500 to 600 µs (read/write respectively) on XtremIO.

**Failover test**

In this test, we simulated a host failover in the following scenarios:

- **Host level**
  - **Planned host failover**—A manual failover used to simulate a planned host failover for system upgrade or patching. The test shows how the solution maintains a continuous working environment on the remote site with the same performance expectation while the production site is undergoing system upgrade or patching.
  - **Unplanned host failure with automatic failover**—Simulated by powering off the production SQL Server to cause a host failure that triggers an automatic failover to the secondary site. The test shows how the solution provides high availability during unexpected server failure on the production site.
• Site level
  • Site failure with automatic failover—Simulated failure where the server, VPLEX, and XtremIO are offline or not reachable on the production site. The test shows how the solution is capable of sustaining the same SQL Server database performance with minimal client disruption after a general site failure.

Test methodology
This test used BenchCraft creating the same workload as in performance testing while measuring the performance of the environment.

The production and secondary sites each maintained a set of snapshots that was created at the same time. A full workload was not running when we measured the baseline performance. The workload on the SQL Server cluster was stopped before the planned failover took place to simulate a well-planned failover. The time duration of the SQL Server cluster failing over was measured and the workload was generated against the same cluster after failover to measure the performance on the secondary site.

A full workload ran on the environment. Failover was triggered automatically after the host was powered off. We measured the time duration of the SQL Server cluster failing over to the secondary site. The performance was measured with the same workload going to the same cluster but running on the secondary site.

To measure the system on the same level, the workload on the snapshot was also simulated with the mount host on the secondary site.

Test procedure
Planned host failover
We used the following steps for the planned host failover test:

1. Generate the workload for the entire environment of the production database including the snapshot workload.
2. Capture the latency, IOPS, and transaction rate as a baseline.
3. Fail over the SQL Server 2012 cluster:
   a. Stop the SQL Server 2012 workloads.
   b. Manually fail over the SQL Server 2012 and measure the duration of the failover.
   c. Reconnect the workload to the SQL Server 2012 cluster, which is now running on the secondary site.
   d. Capture all performance counters with the full workload.
4. Fail over the SQL Server 2014 cluster:
   a. Stop the SQL Server 2014 workloads.
   b. Manually fail over the SQL Server 2014 and measure the duration of the failover.
c. Reconnect the workload to the SQL Server 2014 cluster, which is now running on the secondary site.

5. Reconnect the workload for the snapshot from the snapshot to the snapshot mounted on the secondary site and measure the performance.

Unplanned host failure with automatic failover

We used the following steps for the unplanned host failure test:

1. Generate the workload for the entire environment of the production database including the snapshot workload.

2. Capture the latency, IOPS, and transaction rate as a baseline.

3. Fail over the SQL Server 2012 cluster:
   a. Simulate an unplanned host failure by powering off the ESXi Server under the full workload.
   b. The system automatically fails over the SQL Server 2012 to the secondary site. Measure the duration of the failover.
   c. Reconnect the workload to the SQL Server 2012 cluster, which is now running on the secondary site.
   d. Capture all performance counters with the full workload.

4. Fail over the SQL Server 2014 cluster:
   a. Simulate an unplanned host failure by powering off the ESXi Server under the full workload.
   b. The system automatically fails over the SQL Server 2014 to the secondary site. Measure the duration of the failover.
   c. Reconnect the workload to the SQL Server 2014 cluster, which is now running on the secondary site.

5. Reconnect the workload for the snapshot from the snapshot to the snapshot mounted on the secondary site and measure the performance.

Site failure with automatic failover

We used the following steps for the site failure test:

1. Generate the workload for the entire environment of the production database including the snapshot workload.

2. Capture the latency, IOPS, and transaction rate as a baseline.

3. Simulate a site failure by powering off the XtremIO X-Bricks server and ESXi Server on the production site.

4. Observe the cluster failover on the secondary site and measure the duration of entire site failover.

5. Reconnect the workload to the SQL Server 2012 and SQL Server 2014 production cluster now running on the secondary site.
6. Start the same workload on the mounted snapshot copies of the secondary site.

7. Verify the system performance.

**Test results**

During all of our failover testing, when we completed the SQL Server failover, the database recovered to full function on the second site in about one minute, even during the total site failure case.

When VPLEX was configured with synchronous consistency groups, data was only committed to the database when the remote site write was completed. This resulted in very little to no recovery time to roll back the uncommitted transactions after failover, as shown in Figure 17.

![Figure 17. Extremely low number of transactions rolled back after failover](image)

**Planned host failover**

Planned host failover is used when there is a need to patch or upgrade the production server. The database administrator schedules a planned failover to the secondary site. This type of failover is usually scheduled during the less busy hours to ensure a smooth system upgrade.

During our testing, the planned failover took about 15 seconds to fail the cluster to the secondary site and recover. The system took about 30 seconds to start accepting client connections again and return to full functionality. All the databases were recovered without any issues and continued running a heavy workload.

As shown in Figure 18, the planned failover took less than one minute to fail over with or without workload. Recovery depended on many environmental variables. Both SQL Server and MSCS cluster took time to prepare the database for the secondary site to take over. The total time it took from initiation of the failover to the time when the database was fully functional for the workload was about 50. The difference was not significant and was most likely due to a fluctuation in the system.
The environment sustained the same performance with the production when all the servers failed over to the secondary site. If only one server failed over to the secondary site, we actually utilized the resources from both sites. You can expect the performance to be much better than when utilizing only one site.

In our testing, when only the SQL Server 2012 cluster was failed over to the secondary site, the total throughput for the whole environment increased about 15 percent. This could improve even more, if we increased the workload. The latency in all instances was dropped because the workload was offloaded from the primary storage.

This environment can support an active/active configuration with each site hosting a SQL Server cluster production, while providing support for high availability through VPLEX virtualized XtremIO storage in a Microsoft multisite cluster.

*Unplanned host failure with automatic failover*

Unplanned host failure can happen at any time. To demonstrate the resilience of the VPLEX encapsulated XtremIO configuration for a SQL Server stretched cluster, we simulated the system with full workload at the time of system failure.

The client workload was reconnected to the cluster immediately after the failure and performance was measured. The SQL Server stretched cluster failed over within about 30 second. SQL Server databases recovered and were fully functional within 54 seconds and continued to run the heavy workload without any issue.

As shown in Figure 19, the cluster took about 3 to 5 seconds to ensure the primary host was down due to an issue and not just a network “hiccup”. The failover took approximately the same amount of time as the planned failover, and the database recovery was processed in about 30 seconds.
Performance impact of a host failure

This solution experience minimal performance impact during a host failure. The VPLEX sustained the same amount of workload after failover.

Overall, when the entire database failed over to the secondary site and the full workload was restored, the database performance was almost the same as it was when the SQL Server clusters were on the primary site; the difference was negligible.

When only one of the clusters failed over, both primary and secondary site resources were utilized, therefore the total IOPS was much higher.

Site failover with automatic failover

When the production site was simulated to create a power failure, the entire production server failed over to utilize the secondary site and continue providing database servicing. Our test showed:

- In a fully loaded database environment, when the production site experienced unexpected site failure, this architecture provided a robust, seamless transition to the secondary site and full functionality without much delay. The observed failover time in our testing was about 1 minute for the whole site.

- When we reconnected the full workload to the environment, the client connected within a minute. Performance was restored to the pre-failover level. The business interruption for a disaster such as site failure caused minimum impact for the end user in both function and performance.

During this test, we also redirected all the test/development and QA workloads to the secondary site after the simulated site failure. The environment sustained the same workload with similar or slightly better than pre-failure performance.

Workload on both the production cluster and snapshots achieved a similar level of performance as the production site as long as the total IOPS capacity was within the limits of XtremIO and VPLEX.
Conclusion

Summary

This solution demonstrates that the EMC XtremIO all-flash array benefits can be sustained across multiple data centers for resiliency and without compromising availability, configuration, or performance. XtremIO with EMC VPLEX offers a scalable, highly available, and extremely efficient storage solution for a consolidated Microsoft SQL Server environment that can be used for various workloads, especially for OLTP. XtremIO virtualized by VPLEX Metro provides highly available storage for SQL Server clusters. XtremIO N-way active/active scale-out architecture linearly scales capacity, support extremely high throughputs, and maintains minimal latency.

The solution demonstrates efficient resource utilization through virtualization while providing high database performance. The capacity and processing power can easily be increased. As business needs change, this solution stack can align with the shifting demands from any level, such as applications, database software, and non-database software. In turn, multiple new workload approaches, such as realtime analytics, are made possible, with the consolidation of production and reporting instances.

Instant failover can be achieved in this highly available environment. Whether the need for a system upgrade or software patch requires that the production host is temporarily out of service, or a hardware failure causes a host-level or even site-level failure, the entire system will recover on the secondary site and continue serving the SQL Server client.

Findings

This solution provides:

- Fast and simple setup with little to no storage tuning. VPLEX virtualized XtremIO works as seamlessly in virtualized SQL Server environments as in physical ones, and is easy to manage and monitor.

- Support for the most demanding transactional Microsoft SQL Server 2012 and Microsoft SQL Server 2014 workloads with high availability built on Microsoft Failover Clustering and VPLEX storage utilization.

- Instant failover, which helps deliver a smoother system upgrade or patch process. All client activities are kept at the same level while the production server is down for upgrade or maintenance.

- In the event of a host or server failure or even an entire site failure, this configuration ensures sustained performance and minimized disruption for the client application.
These documents are available from the EMC.com or EMC Online Support websites. Access to online support depends on your login credentials. If you do not have access to a document, contact your EMC representative.

For additional information, refer to the following white papers:

- *Introduction to the EMC XtremIO All-Flash Array*
- *Microsoft Windows Server Failover Clustering with EMC VPLEX—Best Practices Planning*
- *Using VMware vSphere with EMC VPLEX—Best Practices Planning*

For additional information, refer to the following product documents:

- *EMC VPLEX Metro Witness Technology and High Availability TechBook*
- *EMC VSI for VMware vSphere Path Management Product Guide*

For more information, refer to the following document on the XtremIO website:

- *EMC XtremIO System Specifications Data Sheet*
- *EMC XtremIO Storage Array User Guide*

For additional information, refer to the following documents or topics on the VMware website:

- *Microsoft SQL Server Databases on VMware Best Practices Guide*
- *VMware vSphere Networking*
- *VMware vSphere ESXi Server*
- *VMware ESX Scalable Storage Performance*
- *Setup for Failover Clustering and Microsoft Cluster Service*

For additional information, refer to the following document on the Microsoft website:

- *Microsoft SQL Server 2014*

For additional information, refer to the following topics on the MSDN website:

- *AlwaysOn Client Connectivity (SQL Server)*
- *Create or Configure an Availability Group Listener (SQL Server)*
- *Pre-Configuration Database Optimizations*