EMC MULTISITE DISASTER RECOVERY FOR MICROSOFT SQL SERVER 2012

EMC VNX5700, EMC FAST Cache, SQL Server AlwaysOn Availability Groups

- Mission-critical high availability
- Automated storage optimization

EMC Solutions Group

Abstract

This white paper describes how to deploy and implement best practices for a geographically dispersed Microsoft SQL Server 2012 environment with AlwaysOn Availability Groups on an EMC® VNX5700™ platform. The white paper also demonstrates how EMC FAST Cache provides significant storage performance improvements for SQL Server 2012 in an automated, nondisruptive fashion.

April 2012
# Table of contents

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

**Introduction** ...................................................................................................................................... 7  
Purpose ........................................................................................................................................... 7  
Scope .............................................................................................................................................. 7  
Audience .......................................................................................................................................... 7  
Terminology ..................................................................................................................................... 7  

**Technology overview** ........................................................................................................................ 10  
Overview ........................................................................................................................................ 10  
EMC VNX5700 ................................................................................................................................ 10  
  EMC Unisphere .......................................................................................................................... 10  
EMC FAST Cache ............................................................................................................................. 10  
EMC Storage Integrator .................................................................................................................. 11  
EMC PowerPath .............................................................................................................................. 11  
Microsoft SQL Server 2012 ............................................................................................................. 11  
  AlwaysOn .................................................................................................................................. 12  
  Availability groups .................................................................................................................... 12  

**Solution configuration** ...................................................................................................................... 13  
Overview ........................................................................................................................................ 13  
Physical environment ....................................................................................................................... 13  
Hardware resources ........................................................................................................................ 14  
Software resources ........................................................................................................................ 14  
Environment profile ........................................................................................................................ 15  

**SQL Server 2012 AlwaysOn** .............................................................................................................. 16  
SQL Server high availability solutions overview .............................................................................. 16  
AlwaysOn failover clustering .......................................................................................................... 16  
AlwaysOn Availability Groups ..................................................................................................... 17  
Availability replicas and roles ........................................................................................................ 17  
  Primary and secondary roles ...................................................................................................... 19  
  Secondary replicas ....................................................................................................................... 19  
  Readable secondary replicas ...................................................................................................... 20  
Backup preferences ........................................................................................................................ 20  
Availability modes .......................................................................................................................... 21  
Availability group listener .............................................................................................................. 21
Findings ......................................................................................................................................... 46

References .................................................................................................................................. 48
White papers ................................................................................................................................. 48
Other documentation .................................................................................................................... 48
Executive summary

Business case

Never before has access to mission-critical data been more important to businesses competing in a rapidly changing global economy. Business analytics, inventory management, project management, customer insight and collaboration, and business functions that rely on the availability of Microsoft applications, demand that applications and data be available 24x7. But it is not enough just to deliver business-class availability; the information needs to be delivered immediately to users in any location, on any device.

Working together, Microsoft and EMC deliver high-performance business-class availability solutions for Microsoft SQL Server 2012 databases and applications. By combining the power of SQL Server 2012 AlwaysOn Availability Groups and EMC® VNX™ unified storage, this joint solution improves database availability and increases overall system utilization.

The EMC VNX series consists of high-performing unified storage with unsurpassed simplicity and efficiency, optimized for physical or virtual applications. VNX5700 can help meet the requirements to reduce cost and complexity while providing the power and flexibility to meet database and application demands, especially when combined with EMC FAST Cache technology, which extends the system memory with Flash drives for realtime performance improvement. Using the VNX platform with FAST Cache, SQL Server 2012 deployments can gain a significant performance boost without the need to manually redesign storage or make changes at the application level.

Solution overview

The solution is based on a geographically dispersed SQL Server 2012 environment protected by AlwaysOn technology, and highlights multi-subnet support at both synchronous and asynchronous distances.

A set of mission-critical, active online transaction processing (OLTP) databases are replicated to the disaster recovery site using the SQL Server AlwaysOn Availability Groups feature.

Key results

Testing and validation demonstrated that:

- SQL Server AlwaysOn Availability Group technology brings significant improvements to the levels of native protection in terms of ease of use, monitoring, and performance compared to legacy native SQL Server protection offerings in previous versions.

- EMC VNX5700 can easily service 50,000+ SQL Server online transaction processing (OLTP-like) IOPS while providing high availability (HA) through the SQL Server 2012 AlwaysOn Availability Groups configuration.

- The results highlight the simplicity of enabling FAST Cache technology to receive significant boosts in SQL Server transactional performance.

Note: The VNX5700 array can handle far more IOPS than recorded in this white paper. The performance achieved in this solution was reached with the given amount of storage devices (refer to the Solution configuration section).
Introduction

Purpose

The solution showcases the AlwaysOn high availability features of AlwaysOn Availability Groups, focusing on the SQL Server 2012 AlwaysOn transaction-level replication technology. This maximizes the availability of a set of user databases between two distant, multi-subnet sites while simultaneously boosting database storage efficiency and optimization using the FAST Cache technology available on the VNX storage series.

Scope

The scope of this white paper is to:

- Validate the VNX storage platform for critical, enterprise-level, SQL Server 2012 deployments in a multi-subnet disaster recovery configuration featuring the SQL Server 2012 AlwaysOn Availability Groups.
- Demonstrate VNX5700's ability to easily service high-scale OLTP workloads for SQL Server 2012 environments. (This solution by no means reached the capability of the VNX5700 in terms of capacity or performance).
- Showcase the ability of EMC FAST Cache technology to significantly boost the performance of SQL Server 2012 deployments.

Audience

This white paper is intended for EMC employees, partners, and customers, including IT planners; storage architects, SQL Server database administrators, and EMC field personnel who are tasked with deploying such a solution in a customer environment. It is assumed that the reader is familiar with the various components of the solution.

Terminology

This paper includes the terminology defined in Table 1.

Table 1. Terminology

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active secondary</td>
<td>Secondary replica databases that are actively used, for example, configured as readable secondary replicas.</td>
</tr>
<tr>
<td>Automatic failover</td>
<td>Automatic failover is a failover option for availability groups in synchronous-commit mode.</td>
</tr>
<tr>
<td>Availability database</td>
<td>A database that belongs to an availability group. For each availability database, the availability group maintains a single read/write copy (the primary database) and one to four read-only copies (secondary databases).</td>
</tr>
<tr>
<td>Availability group</td>
<td>A container for a set of availability databases that fail over together.</td>
</tr>
<tr>
<td>Availability group listener</td>
<td>A server name to which clients can connect in order to access a database in a primary or secondary replica of an AlwaysOn Availability Group. Availability group listeners direct incoming connections to the primary replica or to a read-only secondary replica.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Availability replica</td>
<td>An instantiation of an availability group that is hosted by a specific instance of SQL Server and maintains a local copy of each availability database that belongs to the availability group. Two types of availability replicas exist: a single primary replica (refer to the Primary replica definition) and one to four secondary replicas (refer to the Secondary replica definition).</td>
</tr>
<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>Failover partners</td>
<td>Availability groups consist of a set of two or more failover partners referred to as availability replicas.</td>
</tr>
<tr>
<td>FAST Cache</td>
<td>Fully Automated Storage Tiering Cache is a performance optimization feature of the EMC VNX series storage arrays. FAST Cache uses Flash drives to extend existing cache capacities for accelerated system performance and automatic absorption of unpredicted spikes in application workloads.</td>
</tr>
<tr>
<td>FAST VP</td>
<td>Fully Automated Storage Tiering for Virtual Pools. FAST VP provides automatic storage tiering at the sub-LUN level.</td>
</tr>
<tr>
<td>Flexible failover policy</td>
<td>The flexible failover policy of an availability group provides you with granular control over what conditions must cause an automatic failover.</td>
</tr>
<tr>
<td>Forced failover</td>
<td>Forced failover is the only option available for availability groups in asynchronous-commit mode.</td>
</tr>
<tr>
<td>IOPS</td>
<td>I/O (input/output) per second. A measure of disk performance in terms of I/O command processing throughput per second.</td>
</tr>
<tr>
<td>LUN</td>
<td>Logical unit number. An identifier used to describe and identify logical storage objects of a storage subsystem.</td>
</tr>
<tr>
<td>Manual failover</td>
<td>Manual failover is a failover option for availability groups in synchronous-commit mode.</td>
</tr>
<tr>
<td>Multi-subnet</td>
<td>A SQL Server multi-subnet failover cluster is a configuration where failover cluster nodes are connected to different networks. These networks can be in the same physical location or in physically dispersed networks.</td>
</tr>
<tr>
<td>OLTP</td>
<td>Online transaction processing (such as a workload from a trading or banking application).</td>
</tr>
<tr>
<td>Primary database</td>
<td>The production read/write copy of an availability database.</td>
</tr>
<tr>
<td>Primary replica</td>
<td>The availability replica that makes the primary databases available for read/write connections from clients and sends transaction log records for each primary database to every secondary replica.</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>RDM</td>
<td>Raw device mapping.</td>
</tr>
<tr>
<td>Readable secondary replica</td>
<td>Secondary replica databases, which are configured to enable read-only client connections (Refer to the Secondary replica definition).</td>
</tr>
<tr>
<td>Reseeding</td>
<td>Process of copying a database from a primary replica to corresponding secondary replicas.</td>
</tr>
<tr>
<td>Secondary database</td>
<td>A read-only copy of an availability database.</td>
</tr>
<tr>
<td>Secondary replica</td>
<td>Each availability group supports one primary replica and up to four secondary replicas. Secondary databases can be made available for read-only access and/or some backup operations.</td>
</tr>
<tr>
<td>Target secondary replica</td>
<td>During failover the target secondary replica transitions to the primary role.</td>
</tr>
<tr>
<td>TPS</td>
<td>Transactions per second</td>
</tr>
</tbody>
</table>
## Technology overview

### Overview

The following components are used in this solution:

- EMC VNX5700
- EMC FAST Cache
- EMC Storage Integrator (ESI)
- EMC PowerPath™
- Microsoft SQL Server 2012
- Microsoft SQL Server 2012 AlwaysOn
- Microsoft SQL Server 2012 AlwaysOn Availability Groups

### EMC VNX5700

EMC VNX5700 is a high-end, enterprise storage array comprising a system bay that includes a storage-processor enclosure, up to two storage processors, with disk array enclosures (DAEs), and separate storage bays that can scale up to 500 disk drives.

VNX5700 arrays support multiple drive technologies, including Flash, serial-attached SCSI (SAS), and nearline SAS (NL-SAS) drives, and the full range of RAID types. The VNX series is powered by Intel Xeon processors, for intelligent storage that automatically and efficiently scales in performance, while ensuring data integrity, security, and 24/7 uptime for mission-critical SQL Server environments.

### EMC Unisphere

EMC Unisphere™ makes it easy to manage VNX systems from anywhere with a simple, integrated user interface for distributed storage environments. The Unisphere dashboard is a single screen for at-a-glance management and reporting, enabling administrators to gain instant and actionable knowledge about the environment. Unisphere’s single sign-on automatically discovers all VNX, EMC CLARiiON™, EMC Celerra™, and EMC RecoverPoint SE installations in the environment for seamless configuration.

### EMC FAST Cache

The VNX series supports an optional performance-enhancing feature called FAST Cache. FAST Cache is a specially-configured set of Flash drives that serves as an I/O accelerator. It is available as a global resource and can support LUNs in RAID groups and pools. With a compatible workload, FAST Cache increases performance in the following ways:

- Reduces response time to hosts
- Enables lower pool and RAID group physical drive utilization

**Note** FAST Cache supports both pool-based and traditional LUNs.
Apart from being able to cache read and write I/Os, the storage processors coalesce writes and pre-fetch reads to improve performance. However, these operations generally do not accelerate random read-heavy I/Os. This is where FAST Cache helps. FAST Cache monitors the storage processors’ I/O activity for blocks that are read or written to multiple times in storage, and promotes those blocks into the FAST Cache if they are not already cached. The blocks can subsequently be accessed at far lower response times.

**EMC Storage Integrator**

EMC Storage Integrator (ESI) is an agentless, no-charge plug-in enabling application-aware storage provisioning for Microsoft Windows server applications. Windows administrators can easily provision (using wizards) block and file storage in a Windows environment. ESI supports the EMC VNX, EMC VNXe, EMC CLARiiON CX4, and EMC VMAX and VMAXe storage platforms.

**EMC PowerPath**

EMC PowerPath software was used on the vSphere host in the VMware HA cluster. PowerPath allows the host to connect to a LUN through more than one storage processor port; this is known as multipathing. PowerPath optimizes multipathed LUNs through load-balancing algorithms. Port-load balancing equalizes the I/O workload over all available channels. Hosts connected to VNXs benefit from multipathing. The advantages of multipathing are:

- Failover from port to port on the same storage processor, maintaining an even system load and minimizing LUN trespassing
- Port-load balancing across storage processor ports and host bus adapters (HBAs)
- Higher bandwidth attachment from host to storage system

**Microsoft SQL Server 2012**

SQL Server 2012 is the latest version of Microsoft’s database management and analysis system for e-commerce, line-of-business, and data warehousing solutions.

SQL Server 2012 includes technologies such as:

- Database Engine
- Data Quality Services (DQS)
- Analysis Services
- Integration Services
- Master Data Services
- Replication
- Reporting Services (SSRS)

This solution concentrates on one of the new and most significant replication features of SQL Server 2012, AlwaysOn, and in particular AlwaysOn Availability Groups.
**AlwaysOn**

SQL Server AlwaysOn refers to the comprehensive HA and disaster recovery solution for SQL Server 2012. AlwaysOn presents new and enhanced capabilities for both specific databases and entire instances, providing flexibility to support various high availability configurations through:

- AlwaysOn Failover Cluster Instances (FCIs)
- AlwaysOn Availability Groups

AlwaysOn Availability Groups is the focus in this solution, but AlwaysOn FCIs are discussed in the SQL Server high availability solutions overview section.

**Availability groups**

AlwaysOn Availability Groups is a HA and DR solution introduced in SQL Server 2012, enabling administrators to maximize availability for one or more user databases. SQL Server instances are configured so that a single or group of primary databases can have up to four secondary databases residing on Windows Server failover cluster (WSFC) nodes. Availability groups can be configured in two availability modes:

- Asynchronous-commit mode
- Synchronous-commit mode
Solution configuration

Overview

This white paper characterizes and validates a VNX5700 storage array supporting a geographically dispersed SQL Server 2012 environment protected by AlwaysOn technology, highlighting multi-subnet support at synchronous and asynchronous distances of 80 km, 800 km, and 4,000 km.

Current VNX storage configuration best practices are used alongside EMC FAST Cache to increase performance for OLTP.

Physical environment

Figure 1 shows the overall physical architecture of the environment.
Table 2 shows the hardware resources used in this solution.

### Table 2. Hardware

<table>
<thead>
<tr>
<th>Equipment</th>
<th>Quantity</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage platform</td>
<td>2</td>
<td>VNX5700</td>
</tr>
<tr>
<td>Fibre switch</td>
<td>2</td>
<td>8 GB 48-port departmental Fibre Channel switches</td>
</tr>
<tr>
<td>FC HBA</td>
<td>4</td>
<td>8 GB (2 per SQL Server production host)</td>
</tr>
<tr>
<td>Network switch</td>
<td>2</td>
<td>1 GB switch 48-port</td>
</tr>
<tr>
<td>Distance simulator</td>
<td>1</td>
<td>IP 1-10 GigE/FC 1-8 GB optic-based physical device</td>
</tr>
<tr>
<td>SQL Server physical servers</td>
<td>2</td>
<td>• 32 core/256 GB memory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Processor: Intel Xeon X7560)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 6 x 1GigE ports</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 2 x dual port 8 GB FC HBAs</td>
</tr>
<tr>
<td>Management and load generators (local and remote sites)</td>
<td>2</td>
<td>• 2 x 12 core/96 GB memory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Processor: Intel Xeon X7560)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Virtualized servers running Hyper-V</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Virtual machines:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 x SQL Server test load server</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 x domain controllers</td>
</tr>
</tbody>
</table>

Table 3 shows the software resources used in this solution.

### Table 3. Software

<table>
<thead>
<tr>
<th>Description</th>
<th>Quantity</th>
<th>Version</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMC VNX Block Operating Environment</td>
<td>1</td>
<td>05.31.000.5.704</td>
<td>VNX operating environment</td>
</tr>
<tr>
<td>EMC Storage Integrator</td>
<td>1</td>
<td>1.3.595.2616</td>
<td>Provisioning new storage and enhanced storage views and reporting</td>
</tr>
<tr>
<td>EMC PowerPath</td>
<td>2</td>
<td>5.5.1</td>
<td>Advanced multipathing for SQL Server production host HBAs</td>
</tr>
<tr>
<td>EMC Unisphere</td>
<td>1</td>
<td>1.1.30.1.0090</td>
<td>VNX management software</td>
</tr>
<tr>
<td>EMC Navisphere® CLI</td>
<td>1</td>
<td>7.31.30.0.90</td>
<td>CLI software to manage the VNX storage array</td>
</tr>
<tr>
<td>Windows Server 2008 R2 Enterprise Edition</td>
<td>9</td>
<td>2008 R2 x64 SP1</td>
<td>Server operating system</td>
</tr>
<tr>
<td>Microsoft SQL Server 2012 Enterprise Edition</td>
<td>2</td>
<td>2012</td>
<td>Database server</td>
</tr>
</tbody>
</table>
This solution was validated with the environment profile listed in Table 4.

Table 4. Environment profile

<table>
<thead>
<tr>
<th>Profile characteristic</th>
<th>Quantity/Type/Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft SQL Server 2012</td>
<td>2 x SQL Server instances</td>
</tr>
<tr>
<td>OLTP Database 1 (OLTP_1)</td>
<td>100,000 users/TPC-E-like/1 TB</td>
</tr>
<tr>
<td>OLTP Database 2 (OLTP_2)</td>
<td>50,000 users/TPC-E-like/500 GB</td>
</tr>
<tr>
<td>OLTP Database 3 (OLTP_3)</td>
<td>25,000 users/TPC-E-like/250 GB</td>
</tr>
<tr>
<td>OLTP Database 4 (OLTP_4)</td>
<td>5,000 users/TPC-E-like/50 GB</td>
</tr>
</tbody>
</table>
**SQL Server 2012 AlwaysOn**

**SQL Server high availability solutions overview**

SQL Server offers administrators several options to configure high availability for both servers and databases. These high availability configurations have until now included:

- **Database mirroring**—maintains a single standby database, or *mirror database*, of a production database that is referred to as a *principal database*. This solution increases database availability by supporting almost instantaneous failover and is primarily a software solution.

- **Log shipping**—maintains one or more warm standby databases, or *secondary databases*, for a single production database that is referred to as the *primary database*.

- **Replication**—a primary server, referred to as the *Publisher*, distributes data to one or more secondary servers, or *Subscribers*. Replication enables real-time availability and scalability across these servers. Replication also supports filtering to provide only a subset of data to subscribers, and also allows for partitioned updates. Subscribers are online and available for reporting and other functions without query recovery. Three types of replication are available: snapshot, transactional, and merge. Transactional replication provides the lowest latency and is the one used most often for high availability.

SQL Server 2012 introduces two high availability configurations as part of SQL Server AlwaysOn, which provides availability at either the application database or instance level:

- AlwaysOn Failover Clustering
- AlwaysOn Availability Groups

**AlwaysOn failover clustering**

A single SQL Server instance is installed across multiple Windows Server Failover Cluster (WSFC) nodes. WSFC functionality provides high availability at the instance level, by presenting an FCI to the network as a single computer accessible through the cluster’s virtual name.

This configuration is an enhancement to the SQL Server FCI functionality available in previous versions of SQL Server. In SQL Server 2008 R2 and previous versions, SQL Server iterated through all IP addresses in the failover cluster resource group and attempted to bind to all during startup, with failed bindings causing SQL Server startup fails. Therefore, in SQL Server 2008 R2 and previous versions, stretch VLANs were used to enable SQL Server multisite failover clustering.

With SQL Server 2012, improvements have been made to the multisite and specifically multi-subnet failover clustering implementation. Two major enhancements made to support multi-subnet clustering are:

- **Cluster Setup support**—Both AddNode, for integrated installation, and CompleteFailoverCluster, for advanced installation, can intelligently detect a multi-subnet environment, and automatically set the IP address resource dependency to OR.
- SQL Server Engine support—To bring the SQL Server resource online, the SQL Server Engine startup logic skips binding to any IP address that is not in an online state, as shown in Figure 2.

**Figure 2. Cluster Properties—Dependencies**

### AlwaysOn Availability Groups

AlwaysOn Availability Groups support a failover environment for a specific set of user databases, known as *availability databases*, which fail over together. Like AlwaysOn Failover Clustering, AlwaysOn Availability Groups require the SQL Server instances to be configured on WSFC nodes, but with the instances remaining and being presented to the network as separate computers. Availability groups support a set of *primary databases* and one to four sets of corresponding *secondary databases*. An availability group fails over at the level of an availability replica and, optionally, secondary databases can be made available for read-only access and some backup operations.

### Availability replicas and roles

Availability groups consist of a set of two or more failover partners referred to as *availability replicas*. Each availability replica is hosted on a separate instance of SQL Server which in turn resides on separate nodes of a WSFC cluster. Each of the SQL Server instances is either a SQL Server FCI, as shown in Figure 3, or a stand-alone instance with AlwaysOn Availability Groups enabled, as shown in Figure 4.
Figure 3. Microsoft SQL Server FCI

Figure 4. Microsoft SQL Server AlwaysOn Availability Group
Table 5 compares FCI nodes and availability group replica features.

### Table 5. Comparing FCI nodes and availability group replicas

<table>
<thead>
<tr>
<th>Feature</th>
<th>Nodes within an FCI</th>
<th>Replicas within an availability group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uses WSFC cluster</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Storage type</td>
<td>Shared</td>
<td>Non-shared</td>
</tr>
<tr>
<td>Performs a failover of storage</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Protection level</td>
<td>Instance</td>
<td>Database</td>
</tr>
<tr>
<td>Number of nodes/replicas</td>
<td>Standard: 2</td>
<td>5 (including primary)</td>
</tr>
<tr>
<td></td>
<td>Enterprise and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Datacenter: 16</td>
<td></td>
</tr>
<tr>
<td>Readable secondary copies</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Applicable failover policy settings</td>
<td>WSFC quorum</td>
<td>WSFC quorum</td>
</tr>
<tr>
<td></td>
<td>FCI-specific</td>
<td>Availability group settings</td>
</tr>
<tr>
<td></td>
<td>Availability group</td>
<td></td>
</tr>
<tr>
<td></td>
<td>settings</td>
<td></td>
</tr>
<tr>
<td>Failed over resources</td>
<td>Server, instance,</td>
<td>Database only</td>
</tr>
<tr>
<td></td>
<td>and database</td>
<td></td>
</tr>
</tbody>
</table>

**Primary and secondary roles**

Each availability replica hosts a copy of the availability databases in the availability group. Each availability replica is assigned an initial role as either the primary or the secondary:

- **Primary replica**—holds the primary role and there can be only one. A primary replica hosts the read/write databases which are known as the primary databases.

- **Secondary replica**—of which there can be up to four, each holding the secondary role, with secondary replicas each hosting read-only databases.

**Secondary replicas**

Each secondary replica has the ability to be configured so that it can transition to become the primary replica during failover.

After the initial availability group has been created, any additional database added to the availability group must be an online, read/write database on the server that hosts the primary replica. Adding a new database to an availability group keeps the database online to clients, but no secondary copy exists until a backup of the data and transaction log has been restored to the secondary replica host (using RESTORE WITH NORECOVERY, which omits the undo phase of the restore process, preserving uncommitted transactions. This allows for restoring additional backups to roll the database further forward in time).
Any new secondary database is marked as being in the RESTORING state until it has finished joining the availability group. Once the secondary database has joined the availability group its state is changed to ONLINE. The availability group then initiates data synchronization between the secondary database and its corresponding primary database.

During synchronization the primary replica sends the transaction log records from the primary database to the secondary replica which in turn writes the transaction log records to disk, referred to as **hardening the log**, and applies the log records to the secondary database.

At times, for example, during a failover, the role of the availability replica may become indeterminate. In this case, the databases within the availability group will be marked as being in a NOT SYNCHRONIZING state and their role will be set to RESOLVING until the role of the replica is resolved. If the availability replica is resolved to the primary role, the databases become the primary databases; if resolved to the secondary role, the databases become secondary databases.

**Readable secondary replicas**

Secondary replicas can be configured so that while in the secondary role they accept read-only client connection to local databases. These secondary databases are referred to as **readable secondary replicas**.

**Note** These secondary databases are not set to read-only. Unlike a read-only database, which is static, a secondary replica, secondary database is dynamic, continuously changing as the corresponding primary database changes are applied to it.

**Backup preferences**

Secondary replicas also support performing copy-only backups of a full database, file, and filegroups, and log backups. During availability group creation you can specify a number of settings to control the selection of the replica used for backups.

Backup preference settings:

- **Prefer Secondary**
  Automated backups for this availability group should occur on a secondary replica. If no secondary replica is available, backups are performed on the primary replica.

- **Secondary only**
  All automated backups for this availability group must occur on a secondary replica.

- **Primary**
  All automated backups for this availability group must occur on the current primary replica.

- **Any Replica**
  Backups can occur on any replica in the availability group.
With these preferences you can also set the backup priority (lowest=1, highest=100) and exclude specific replicas.

**Availability modes**

Each availability group has an availability mode setting. This determines whether the primary replica has to wait for a transaction to be committed on a database until the corresponding secondary replica has written the transaction log to disk (hardening the log).

AlwaysOn Availability Groups support two availability modes:

- **Asynchronous-commit mode**
  
  In asynchronous-commit mode the primary replica commits a transaction, without acknowledgment that an asynchronous-commit replica has hardened the log. Asynchronous-commit mode minimizes transaction latency, allowing the secondary databases to lag behind the primary; this makes data loss possible.

- **Synchronous-commit mode**
  
  In synchronous-commit mode, the primary replica waits for acknowledgment that a synchronous-commit secondary has hardened the log before committing a transaction. Synchronous-commit mode increases transaction latency but protects against data loss; this means that as long as the secondary databases are in a synchronized state with the primary database, committed transactions are fully protected.

**Availability group listener**

Client connectivity can be provided on a database of a specified availability group by creation of an availability group listener. This is a unique DNS name that serves as a virtual network name (VNN) and provides a set of resources that are attached to an availability group to direct client connections to the appropriate availability replica. The availability group listener effectively ensures client connectivity in a multi-subnet environment.

Elements of an availability group listener include:

- Virtual network name (VNN)
- The listener port (listens for incoming requests)
- One or more virtual IPs (VIPs), configured for one or more subnets
- Configured to use Dynamic Host Configuration Protocol (DHCP) or a static IP (static IP is recommended)

The listener directs read/write connections to a primary replica, and read-only connections to an appropriately configured read-only secondary replica. If the primary replica goes offline and the primary role transitions to a target secondary replica, that replica becomes the primary replica and the availability group listener redirects client connections to the new primary. Also, any new read-only connections to the new primary are redirected to a suitably configured secondary, if available.
TCP port 1433, the default port for SQL Server, can be used for simplicity when configuring the availability group listener port. If using an alternate port then the port number needs to be in the client connection string and a suitable firewall setting applied. Also, ensure no other service on the cluster node uses this port to avoid a port conflict occurring.

The availability group listener is also able to share the same port being configured among multiple availability group listeners on one SQL Server instance. However, multiple instances of SQL Server (side-by-side) should not be configured to listen on the same port.

Note  For more information about configuring connection strings for client access to databases protected by availability groups, refer to msdn.microsoft.com.

**Failover modes**

With SQL Server 2012 Availability Groups, failover involves a target secondary replica transitioning to the primary replica and taking over the primary role. Three forms of failover exist. Two of these—automatic failover and planned manual failover—require the replicas to be in synchronous-commit mode but have the potential for no data loss. The third option—forced failover—risks data loss, but is the only option available with asynchronous-commit mode.

- **Automatic failover** (without data loss)
  Automatic failover requires that the primary replica and target secondary replica are running in synchronous-commit mode and the secondary replica is in a synchronized state. Also the WSFC quorum should be configured and meet the conditions specified by the availability groups flexible failover policy. For more information refer to WSFC quorum modes and voting configuration.
  
  Failover occurs in response to a failure of the primary replica. A secondary replica is transitioned to the primary role, and becomes the primary replica. After the original primary replica is recovered and becomes available, it is transitioned to the secondary role. It will not be available to become the primary replica until it is once again in a synchronized state.

  Note  SQL Server FCI do not support AlwaysOn automatic failover; an availability replica hosted by an FCI can only be configured for manual failover. If automatic failover is required the SQL Server instances should be configured as part of a WSFC.

- **Planned manual failover** (without data loss)
  Manual failover requires that the primary replica and target secondary replica are running in synchronous-commit mode and the secondary replica is in a synchronized state.
  
  Failover occurs after a manual failover command is issued causing the target secondary replica to be transitioned to the primary role and become the primary replica. The original primary replica now becomes the secondary replica.

- **Forced manual failover** (possible data loss)
  Forced manual failover is the only option available when the primary replica and secondary replica are running in asynchronous-commit mode. Failover
occurs after a manual failover command is issued causing the target secondary replica to be transitioned to the primary role, and become the primary replica.

**Note**  If primary and secondary replicas are running in synchronous-commit mode and the secondary replica is not in a synchronized state, then forced manual failover is the only option available. As forced manual failover requires manual initiation, it is also a kind of manual failover.

Failovers are not caused by database issues such as a database becoming suspect due to a loss of a data file, deletion of a database, or corruption of a transaction log.

**Administration and monitoring**

The introduction of AlwaysOn Availability Groups presents SQL Server administrators with a range of features and metrics to administrate and monitor. Microsoft has provided a series of out-of-the-box features to enable this, which include:

- SQL Server Management Studio (SSMS)
- System Monitor (Performance Counters)
- Transact-SQL (T-SQL)
- PowerShell

**SQL Server Management Studio**

SSMS provides administrator with the ability to create, edit, and monitor availability groups through the regular Object Explorer window.

Once an availability group is created, a centralized availability group dashboard is available. Right-clicking the AlwaysOn High Availability or availability groups objects in Object Explorer shows details of all configured availability groups and lists the primary instance for the availability group, its failover mode, and any current warnings, as shown in Figure 5.

You can right-click an availability group name in Object Explorer, or click the availability group name in the dashboard to open a detailed dashboard view of the selected availability group, as shown in Figure 5.
Figure 5. Detailed dashboard view of availability group OLTP_AG1

This provides further details such as synchronization state, failover readiness, and options such as a Start Failover Wizard, as shown in Figure 5.

Note If the availability group is running in Asynchronous Commit mode the secondary replica is always shown as Synchronizing, as lag occurs and the secondary remains behind.

The dashboard is also highly configurable. It allows you to edit the default view and show numerous additional settings and statistics. For more information on customizing the dashboard view refer to msdn.microsoft.com.
## AlwaysOn Availability Group configuration

**Overview**
The following sections outline the process for implementing AlwaysOn Availability Groups across two SQL Server instances. The primary server was located in the production site and the secondary server in the DR site spanning multiple subnets. Both servers were connected to two VNX5700s, one per site.

**Prerequisites**
Prior to implementing AlwaysOn Availability Groups, always check the latest requirements. Current prerequisites are listed on [msdn.microsoft.com](http://msdn.microsoft.com).

**WSFC**
As part of the configuration for SQL Server 2012 availability group creation, we configured a two-node multisite cluster on Windows Server 2008 R2. More information on WSFC and SQL Server AlwaysOn Availability Groups setup is available on [msdn.microsoft.com](http://msdn.microsoft.com).

**WSFC quorum modes and voting configuration**
SQL Server AlwaysOn Availability Groups and AlwaysOn FCIs both take advantage of WSFC technology. WSFC uses a quorum-based approach to monitoring overall cluster health and provide maximum node-level fault tolerance.

Every node in a WSFC cluster issues heartbeat signals in order to share the status of the node’s health with other nodes in the cluster. Nodes that fail to respond are considered failed. A quorum node set is a majority of the voting nodes and witnesses in a WSFC cluster. The status and health of a WSFC cluster is based on a periodic quorum vote. The presence of a quorum means that the cluster is in a healthy state and can provide node-level fault tolerance.

If a quorum vote fails then the WSFC cluster will go into an offline state and cause all SQL Server instances in the cluster to stop. It is important to note that if the WSFC cluster goes offline due to a quorum failure then manual intervention is required to start the WSFC cluster again.

With Windows Server 2008 R2 four types of quorum modes are available:

- **Node Majority**—Used when you have an odd number of hosts where each node gets one vote. In this configuration there is no file share or disk witness. Over half the nodes must be available to make a quorum.

- **Node and File Share Majority**—Used in a configuration with an even number of nodes in a multisite cluster where a file share is configured to be a voting witness. This allows for the loss of half the nodes if the disk witness is available. If the disk witness is not available, over half the nodes must be available to make a quorum.

- **Node and Disk Majority**—Used in a configuration with an even number of nodes but not in a multisite cluster. In this configuration nodes and a shared disk get voting rights. This allows for the loss of half the nodes if the disk witness is available. If the disk witness is not available, over half the nodes must be available to make a quorum.

- **Disk Only**—In this configuration a shared disk resource is used as the witness. The disk witness needs to be available to have a quorum.
As this solution has a multisite failover cluster with an even number of nodes, the quorum configuration of **Node and File Share Majority** was used. A file share witness was configured on a host outside of the cluster. In this solution a share called Witness was created on a server named ProdDC. The cluster named **CorkCluster** was given read/write permissions to the share at both share-level and NTFS-level permissions.

After configuring the shared folder with the appropriate permissions, we then changed the quorum type. This was done through Failover Cluster Manager, by following these steps:

1. Right-click the cluster, select More Actions and then Configure Cluster Quorum Settings.
2. Select **Node and File Share Majority** as the quorum configuration.
3. Follow the wizard to completion, entering the path to the created file share called Witness.

After completing the remaining steps in the wizard, the Quorum Configuration showed **Node and File Share Majority** in the Failover Cluster Manager, as shown in Figure 6.

**Figure 6.** Failover Cluster Manager—Quorum Configuration

One of the major considerations for node and file share majority quorum is where to place the file share witness. The three options for placement are:

- **File share in primary site**
  
  If there is a failure of the entire production site then the secondary node in the DR site will not come online automatically and will require user input to manually force the quorum online, as this will be the only remaining vote in the cluster.

- **File share in secondary site**
  
  While this eliminates the issue of automatic recovery on the DR site it may also result in false failovers. For example if the DR site fails the primary will also go to an offline state as it no longer has connectivity to the witness and is therefore a single vote scenario.
- File share in a third site

This is the best option for the placement of the file share quorum. In the event of an entire site loss of production or DR the file share on a third site prevents any one site from being a point of failure. In this scenario the cluster will automatically failover in the event of a site loss.

In this solution we had a two-site scenario and therefore implemented the file share in primary site option and placed the file share majority quorum on a production site host.

To enable AlwaysOn high availability in SQL Server Configuration Manager:

1. Open SQL Server Configuration Manager.
2. Right click SQL Server engine services.
3. Click **AlwaysOn High Availability** and select the **Enable AlwaysOn Availability Groups** checkbox, as shown in Figure 7. You must do this for both instances on the two nodes of the cluster.

**Note** Enabling AlwaysOn requires SQL Server services to be restarted on both the nodes and instances.

To enable AlwaysOn high availability in SQL Server Configuration Manager:

1. Open SQL Server Configuration Manager.
2. Right click SQL Server engine services.
3. Click **AlwaysOn High Availability** and select the **Enable AlwaysOn Availability Groups** checkbox, as shown in Figure 7. You must do this for both instances on the two nodes of the cluster.

**Note** Enabling AlwaysOn requires SQL Server services to be restarted on both the nodes and instances.

**Figure 7.** Enable AlwaysOn High Availability in SQL Server Configuration Manager
Storage design

Overview

With SQL Server AlwaysOn availability groups that are replicating primary databases on a production site to secondary replica copies on a DR site, it is important that you consider storage performance after failover to the DR site.

In the following section, consideration is given to the sizing of the storage pool for hosting SQL Server data files. In this solution transaction logs and Tempdb files are segregated to dedicated spindles, hosting traditional RAID 1/0 RAID groups.

Consideration is also given to the impact of FAST Cache has in significantly reducing the volume of mechanical spinning disks required by VNX storage arrays to service the target workload.

Sizing VP pool and FAST Cache

This section looks at the impact of FAST Cache when determining the required number of spinning disk to service a given workload. For this solution we used a given OLTP-like workload generating 50,000 IOPS, with a read/write ratio of approximately 9:1. Due to the impact of write penalties in RAID configurations there are larger back-end IOPS for RAID 5 configurations compared to RAID 1/0, each configuration having its own advantages regarding given workload profiles.

Note: The potential increase in performance provided by FAST Cache is dependent on the workload and the configured cache capacity. Workloads with high locality of reference, small block size, random read I/O, and high concurrency benefit the most from FAST Cache. Workloads made up of sequential I/Os benefit the least.

A RAID 5 protection level was used for the storage pool, giving a balance between capacity and the capability to service the 9:1 read/write ratio workload. The storage pool was created as a homogeneous pool with 40 x 2.5” 10k SAS drives. A FAST Cache configuration of 14 x 3.5” 100 GB Flash drives balanced the solution capability, allowing 641 GB of FAST Cache to be available.

In this solution the 40 x SAS disk pool was able to service 7,652 host IOPS at 65 percent utilization while FAST Cache serviced 42,348 host IOPs.

Table 6 demonstrates the number of physical spinning mechanical drives required to service a 50,000 I/O workload, with the use of FAST Cache to boost performance capabilities and significantly reduce disk numbers.

<table>
<thead>
<tr>
<th>Drive Type</th>
<th>No of Disks</th>
<th>Specifications</th>
<th>RAID Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash drives</td>
<td>14</td>
<td>100 GB</td>
<td>RAID 1</td>
</tr>
<tr>
<td>SAS (10k)</td>
<td>40</td>
<td>600 GB</td>
<td>RAID 5</td>
</tr>
</tbody>
</table>
Calculate the back-end disk IOPS
The host IOPS, as shown in Table 6, is 50,000. The back-end IOPS count is typically larger than the host IOPS due to varying I/O write penalties, depending on the RAID protection technology used.

To determine the amount of drives required for a SAS-only homogenous pool to cater for 50,000 IOPS, we calculated the corresponding back-end IOPS for a RAID 5 pool, as follows:

Syntax:
Back-end disk IOPS = Host read IOPS + 4 x Host write IOPS

Calculation:
Back-end disk IOPS = (0.9 x 50,000 + 4 * (0.1 x 50,000)) = 65,000

Calculate the disk requirements for the SAS-only configuration
In this solution, 600 GB 10k 2.5” SAS disks (150 IOPS per disk capability) were used for the pool. For comparison, the following calculations include the disk requirements for these disks and for 300 GB 15k 3.5” SAS disks, which can perform at 180 IOPS for small block random I/O performance (as is the I/O workload for this environment).

Syntax:
Disk requirement = Back-end disk IOPS / IOPS per disk capability

Calculation (600 GB 10k SAS drives)
65,000 / 150 = 433 disks

Calculation (300 GB 15k SAS drives)
65,000 / 180 = 361 disks

FAST Cache and SAS-only solutions compare
The FAST Cache configuration for the solution was created with 14 x 100 GB Flash drives and 40 x 600 GB 10k SAS drives. Table 7 shows the number of SAS drives without FAST Cache that is required to service the 50,000 I/O workload. The table compares the 300 GB 15K SAS drive and 600 GB 10K SAS drive configurations to the FAST Cache solution.

<table>
<thead>
<tr>
<th></th>
<th>FAST Cache solution</th>
<th>300 GB 15K SAS</th>
<th>600 GB 10K SAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flash drives</td>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SAS drives</td>
<td>40</td>
<td>361</td>
<td>433</td>
</tr>
<tr>
<td>Total Front-End IOPS</td>
<td>50,000</td>
<td>50,000</td>
<td>50,000</td>
</tr>
</tbody>
</table>

Figure 8 illustrates how 14 Flash drives and 40 SAS disks providing 17.15 TB of usable storage is a better fit for 1.8 TB of databases with ample room for growth. The equivalents of SAS-only pools for both 300 GB 15k and 600 GB 10K:

- 433 x 600 GB 10k SAS disks provide a usable capacity of 185.65 TB of storage. This leaves around 183.85 TB of unused storage
- 361 x 300 GB 15k SAS disks provide a usable capacity of 77.39 TB of storage. This leaves around 75.59 TB of unused storage
This clearly shows how the proper use of FAST Cache technology enables the VNX5700 to be used to its optimal performance level and provides greater return on investment (ROI) and reduced total cost of ownership (TCO).

The FAST Cache drives and 40 SAS perform to the stated 50,000 IOPS, which again shows how the 14 Flash drives are an excellent fit for an intensive OLTP-type workload.

The production storage configuration for the solution is shown in Table 8.

Table 8. Production array storage configuration

<table>
<thead>
<tr>
<th>RAID type</th>
<th>Pool/RAID group</th>
<th>Disk configuration</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAID 5</td>
<td>SAS drive pool (Pool—Production OLTP data)</td>
<td>40 x 2.5” 600 GB 10k SAS</td>
<td>OLTP database data files</td>
</tr>
<tr>
<td>RAID 5</td>
<td>SAS RAID group (RAID group 0)</td>
<td>4 x 3.5” 600 GB 10k SAS</td>
<td>SQL Server system databases</td>
</tr>
<tr>
<td>RAID 1/0</td>
<td>SAS RAID group (RAID group 1)</td>
<td>4 x 2.5” 600 GB 15k SAS</td>
<td>SQL Server TempDBs</td>
</tr>
<tr>
<td>RAID 1/0</td>
<td>SAS RAID group (RAID group 2)</td>
<td>8 x 2.5” 600 GB 10k SAS</td>
<td>OLTP database logs</td>
</tr>
<tr>
<td>RAID type</td>
<td>Pool/RAID group</td>
<td>Disk configuration</td>
<td>Purpose</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------</td>
<td>------------------------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>RAID 1</td>
<td>Flash drives</td>
<td>14 x 3.5” 100 GB Flash drives</td>
<td>FAST Cache pool</td>
</tr>
</tbody>
</table>

The DR storage configuration for the solution is shown in Table 9.

**Table 9. DR array storage configuration**

<table>
<thead>
<tr>
<th>RAID type</th>
<th>Pool/RAID group</th>
<th>Disk configuration</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAID 5</td>
<td>SAS drive pool (Pool – Production OLTP Data)</td>
<td>40 x 2.5” 600 GB 10k SAS</td>
<td>OLTP databases data files</td>
</tr>
<tr>
<td>RAID 5</td>
<td>SAS RAID group (RAID group 0)</td>
<td>4 x 3.5” 600 GB 10k SAS</td>
<td>SQL Server system databases</td>
</tr>
<tr>
<td>RAID 1/0</td>
<td>SAS RAID group (RAID group 1)</td>
<td>4 x 3.5” 600 GB 15k SAS</td>
<td>SQL Server TempDBs</td>
</tr>
<tr>
<td>RAID 1/0</td>
<td>SAS RAID group (RAID group 2)</td>
<td>8 x 2.5” 600 GB 10k SAS</td>
<td>OLTP database logs</td>
</tr>
</tbody>
</table>
Table 10 details the storage LUNs provisioned for the solution from the VNX5700 array.

**Table 10. SQL Server LUN design**

<table>
<thead>
<tr>
<th>Name</th>
<th>RAID Type</th>
<th>Storage Pool</th>
<th>User Capacity (GB)</th>
<th>Current Owner</th>
<th>FAST Cache</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP1_DATA_1</td>
<td>RAID 5</td>
<td>Pool PROD OLTP Data</td>
<td>450</td>
<td>SP A</td>
<td>On</td>
</tr>
<tr>
<td>OLTP1_DATA_2</td>
<td></td>
<td></td>
<td>450</td>
<td>SP A</td>
<td>On</td>
</tr>
<tr>
<td>OLTP1_DATA_3</td>
<td></td>
<td></td>
<td>450</td>
<td>SP A</td>
<td>On</td>
</tr>
<tr>
<td>OLTP2_DATA_1</td>
<td></td>
<td></td>
<td>200</td>
<td>SP B</td>
<td>On</td>
</tr>
<tr>
<td>OLTP2_DATA_2</td>
<td></td>
<td></td>
<td>200</td>
<td>SP B</td>
<td>On</td>
</tr>
<tr>
<td>OLTP2_DATA_3</td>
<td></td>
<td></td>
<td>200</td>
<td>SP B</td>
<td>On</td>
</tr>
<tr>
<td>OLTP3_DATA_1</td>
<td></td>
<td></td>
<td>175</td>
<td>SP B</td>
<td>On</td>
</tr>
<tr>
<td>OLTP3_DATA_2</td>
<td></td>
<td></td>
<td>175</td>
<td>SP B</td>
<td>On</td>
</tr>
<tr>
<td>OLTP4_DATA_1</td>
<td></td>
<td></td>
<td>100</td>
<td>SP A</td>
<td>On</td>
</tr>
<tr>
<td>OLTP1_Log</td>
<td>RAID 1/0</td>
<td>RAID Group 2</td>
<td>200</td>
<td>SP A</td>
<td>Off</td>
</tr>
<tr>
<td>OLTP2_Log</td>
<td></td>
<td>RAID Group 2</td>
<td>100</td>
<td>SP B</td>
<td>Off</td>
</tr>
<tr>
<td>OLTP3_Log</td>
<td></td>
<td>RAID Group 2</td>
<td>100</td>
<td>SP B</td>
<td>Off</td>
</tr>
<tr>
<td>OLTP4_Log</td>
<td></td>
<td>RAID Group 2</td>
<td>100</td>
<td>SP A</td>
<td>Off</td>
</tr>
<tr>
<td>SYSTEM_DBs</td>
<td>RAID 5</td>
<td>RAID Group 0</td>
<td>10</td>
<td>SP B</td>
<td>Off</td>
</tr>
<tr>
<td>TempDB_Data_1</td>
<td>RAID 1/0</td>
<td>RAID Group 1</td>
<td>400</td>
<td>SP B</td>
<td>Off</td>
</tr>
<tr>
<td>TempDB_Data_2</td>
<td>RAID 1/0</td>
<td>RAID Group 1</td>
<td>400</td>
<td>SP B</td>
<td>Off</td>
</tr>
<tr>
<td>TempDB_Log</td>
<td></td>
<td></td>
<td>200</td>
<td>SP B</td>
<td>Off</td>
</tr>
</tbody>
</table>

Table 10 shows that balancing across the storage processors was performed during LUN provisioning in anticipation of the workload’s requirements.

**Note**  The total I/O being handled by the assigned LUNs is used in balancing, not the number of LUNs assigned to each storage processor.

**Note**  Take note of the ability to separate Flash drives across physical ports on the VNX. For details refer to the *VNX Best Practice* document, especially the *Back-end SAS Port Balancing* section, which is available on EMC online support.

By having a larger number of files, you can increase the number of physical I/O operations that SQL Server can push to the disk at any one time. The more I/O that SQL Server can push down to the disk level, the faster the database will run. This is why we have a number of physical files for each database.
As can be seen from the database file design, data files should be of equal size for each OLTP database. This is because SQL Server uses a proportional fill algorithm that favors allocations in files with more free space.

**Using RAID groups**

The best practice for SQL Server log files is to use RAID 1/0. OLTP logs have a sequential write I/O pattern of 64 KB. Therefore, the configuration of RAID group 2 with 8 x 2.5” 600 GB 10k SAS was best suited for the log file location. Also, it is best practice to isolate the log from data at the physical disk level.

Performance may also benefit if Tempdb is placed on RAID 1/0 configuration. Because Tempdb has a high write rate RAID1/0 is the best configuration to use. In this solution, Tempdb resided on its own dedicated physical disks in a RAID 10 4 x 2.5” 600 GB group.

**Storage pool**

The virtually provisioned pool was easily created and required only these user inputs:

- Pool name
- Disks: number and type
- Protection level: RAID 5, 6, or 1/0

In this solution, RAID 5 protection level was used, as shown in Figure 9. This was created as a homogeneous pool with 40 SAS drives. With 40 drives for a RAID 5 pool, Virtual Provisioning™ creates eight five-drive (4+1) RAID groups.

---

**Figure 9. Storage Pool Properties—General**
A shared storage space is required when performing a Full Initial Data Synchronization as part of availability group creation. In order to minimize the duration of the seeding and reseeding process, users should consider the storage space used for these database backups and restorations.

The backup process has high bandwidth requirements for storage because it is a sequential write/read workload. RAID 1/0 would best suit the seeding/reseeding of databases for AlwaysOn Availability Group creation. However, as adequate space should be reserved to accommodate reseeding of all databases covered by availability groups, an administrator may choose to use other RAID types to accommodate capacity usage.

FAST Cache uses RAID 1 paired drive provisioning to provide both read and write caching, in addition to mirrored data protection. All of the FAST Cache drives must be the same capacity.

FAST Cache is secure. It is built on Unified LUN technology so the data in FAST Cache is as secure as in any other LUN. It is nonvolatile storage that survives power and SP failures. Also, it does not have to re-warm after a power outage. FAST Cache can benefit any LUN—it is controlled at the LUN level for metaLUN and FLARE® LUNs, and at the pool level for pool LUNs.

When practical, EMC recommends that at least four Flash drives be used in a FAST Cache. With a larger number of drives, concurrency is improved and storage processor contention is reduced, resulting in greater efficiency in the caching role.

In this solution, we used 14 x 3.5" 100 GB Flash drives in the FAST Cache configuration, which consist of seven mirrored pairs. We enabled FAST Cache for the homogenous 40 x 2.5" 600 GB 10k SAS disk pool which contained the OLTP databases data files. To distribute the FAST Cache across the buses, we configured eight drives to use Bus 0 and six drives to use Bus 2.

FAST Cache was not enabled for RAID group LUNs hosting log files or TempDB as there is no gain in performance for small sequential workloads and they fill the tracking map with unnecessary entries. As the sequential data is rarely re-used it has no advantage from Fast Cache. EMC recommend not using FAST Cache with workloads characterized by small sequential I/O (less than 128 KB sequential)—reads or writes.

Enabling FAST Cache

This simple, one-click, transparent process requires no modification at the SQL Server level. Creating a FAST Cache disables the storage system’s read/write cache until the process is complete. As a result, the storage system’s write cache is flushed in order to zero it and it is then automatically reconfigured with less memory. While the read/write cache is disabled, overall performance can be adversely affected.

The time it takes to fully configure the FAST Cache depends on the cache’s size and any workload activity on the storage system. Larger FAST Caches take longer to configure. On a quiet system with low activity and small FAST Caches, the configuration can take several minutes. Configuring a large FAST Cache on a loaded storage system may take longer than an hour.
The order in which drives are added to FAST Cache is the order in which they are bound. This is most easily done using `naviseccli`. The switches used for creating, destroying and displaying FAST Cache configurations are:

- `Cache -fast -create`
- `Cache -fast -destroy`
- `Cache -fast -info`

For example, the following command would bind disks 4 and 5 as primary and secondary respectively, and disks 6 and 7 in Bus 0, Enclosure 0:

```
naviseccli -h IP cache -fast -create -disks 0_0_4 0_0_5 0_0_6 0_0_7 -mode rw -rtype r_1
```

Once FAST Cache is created and fully initialized, it is a simple task to enable FAST Cache for the storage pool, as shown in Figure 10.

![Figure 10. Storage Pool Properties—Advanced](image)

As well as administrating storage though the Unisphere management suite, users now have the option to use ESI. ESI can greatly simplify managing, viewing, and provisioning EMC storage in a Windows environment. Moreover, ESI is free to EMC customers.

ESI for Windows provides storage viewing and provisioning capabilities. As part of the viewing capability, it depicts Windows to storage mapping of resources. For storage provisioning, ESI simplifies the steps involved in creating a LUN and preparing the LUN for the steps of partitioning, formatting, and creating a drive letter. ESI also enables the user to create a file share and mount that file share as a network-attached drive in a Windows environment.

ESI supports VMAX, VNX, VNXe, CX4 and NS series for both block and file support. It is also an agentless architecture so administrators do not need to install an agent on the hosts. In this solution, we installed ESI on a Windows host and used it to provision the storage for the SQL Server 2012 hosts in both the production and DR sites. Figure 11 shows the SQL Server hosts and the VNX5700s across the environment. It also shows the storage pools and all LUNs on the production VNX5700 from ESI.
Creating a disk

Creating a disk for a host can easily be done with the Create Disk wizard. The wizard simplifies the various steps involved in creating a LUN in a specific storage pool, mapping a LUN to the Windows host and adding the storage into the host.

For detailed steps on how to use ESI, refer to *Storage Integrator for Windows 1.3 Product Guide*.

**Note** This white paper describes the features available in the current ESI version at the time of this solution validation. EMC constantly improves and updates its products and technology with new features and functionalities. For the latest features and updates, visit [www.EMC.com](http://www.EMC.com) or EMC online support.
Validation

Test objectives
The testing of this solution validated the ability of the VNX5700 storage array to support SQL Server 2012 in instances running OLTP-like workloads that generated over 50,000 IOPS. Tests involved:

- Introducing Flash drives to the storage array and utilizing them with EMC FAST Cache to boost performance.
- Comparing AlwaysOn Availability Groups in the following availability modes:
  - Synchronous-commit mode—Automatic Failover
  - Synchronous-commit mode—Manual Failover
  - Asynchronous-commit mode—Forced Failover

Note: Benchmark results are highly dependent on 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.

Testing methodology
The testing methodology required OLTP-like workloads to be run against the target databases.

Note: The ability of real-world applications to handle loss of connection will vary, based on design. The tool used in testing to generate workload had a specific behavior, which may not be indicative of customer environments.

Test scenarios
EMC used a number of scenarios to test the solution. These included:

- Baseline testing on an 40 x SAS-only storage pool
- Performance testing on a storage pool with FAST Cache enabled
- Comparing and profiling reseeding times for AlwaysOn Availability Groups for the following availability modes:
  - Synchronous-commit mode—Automatic failover
  - Synchronous-commit mode—Planned manual failover
  - Asynchronous-commit—Forced manual failover
- Validate and profile failover for:
  - Manual failover
  - Automatic failover
  - Site disconnection
Performance test procedures

Testing was conducted by running concurrent OLTP-like workloads against the target databases on the WSFC and standalone SQL Server instances.

1. Each SQL Server instance had its own RAID 5 storage pools for data files with transactional logs running on traditional RAID 1/0 groups.

2. A steady state was reached and a baseline performance measurement observed.

3. FAST Cache was enabled with load running and the performance again monitored with a peak performance level being reached.

Note

The workload profiles parameters were not changed during testing. A profile was set to push the utilization of the SAS-only storage pool and show the impact on performance with the introduction of EMC Flash Cache to the environment.

This approach mimics the potential performance impact of enabling FAST Cache on busy OLTP production environments.

It must be noted that FAST Cache is a shared resource and should be sized as appropriate to the maximum workload it is expected to handle.

Test results

Testing was broken down into the following areas:

- Throughput
- Failover

Throughput testing

The following are the key metrics:

- Throughput in IOPS (transfers/sec)
- SQL Server CPU utilization
- Throughput in transactions/sec (TPS)
- Latency
- Physical disk utilization (percent)
- Storage processor utilization (percent)
- Availability group creation times
Throughput in IOPS (transfers/sec)

Throughput was measured using the Microsoft Performance Monitor (perfmon) counter: LogicalDisk – Avg. Disk Transfer/sec.

**Figure 12.** Avg. Disk Transfer/sec. (IOPS) for both primary and secondary replicas

During baseline testing with 40 SAS disks in the storage pool, transactional I/O throughput on the primary replica produced approximately 11,500 IOPS and the secondary replica produced 1,400 IOPS.

After 30 minutes with FAST Cache enabled on the storage pool, we saw an immediate effect on performance. I/O throughput increased to over 19,000 IOPS on the primary replica and to 2,300 on the secondary. After just two hours of FAST Cache running, we saw throughput increase to over 50,000 IOPS on the primary replica.

As an example of the read/writes being replicated between the primary and secondary replicas, perfmon counters were analyzed for a point in time during the FAST Cache steady state for synchronous-commit mode at 80 km for the 1 TB OLTP_1 database, as shown in Figure 13.

**Figure 13.** Comparison of disk reads/writes on primary and secondary replicas
It can be seen that only 3.7 percent of the primary replica read activity occurs on the secondary replica, compared to 89.51 percent of the write activity. During this period, transactions/sec (TPS) for both primary and secondary replicas was 119. This highlights that, with no read access on the secondary replica, the major activity on the secondary is the writes being replicated.

**SQL Server CPU utilization**

As shown in Figure 14, negligible impact was seen when synchronous-commit mode was used up to 80 km. A small rise of 4 percent CPU percentage utilization was seen using asynchronous-commit mode up to 4,000 km. In all synchronization states it can be seen that minimal CPU utilization occurs on the secondary replica when no additional activity is occurring on the secondary replica databases.

![SQL Server CPU impact across synchronization state and distance](image)

**Figure 14. SQL Server CPU impact across synchronization state and distance**

**Throughput in transactions/sec (TPS)**

Throughput was also measured using the Microsoft Performance Monitor (perfmon) counter: Databases – Disk Transactions/sec.

![Disk transactions per sec (TPS) for both primary and secondary replicas](image)

**Figure 15. Disk transactions per sec (TPS) for both primary and secondary replicas**

During baseline testing with 40 SAS disks in the storage pool, transactional activity, shown as SQL Server TPS, was over 4,900 TPS with the secondary replica at 300 TPS.
After 30 minutes with FAST Cache enabled on the storage pool, SQL Server TPS on the primary replica increased to approximately 12,000 TPS and to 900 TPS on the secondary replica. After just two hours of FAST Cache running we saw transactional performance increase to over 24,000 TPS.

Testing was conducted with distance emulation between production and DR sites initially set to 80 km distance during FAST Cache warm-up, with replication set to synchronous-commit mode. Distance was then stretched to 800 km and then 4,000 km, both with replication set to asynchronous-commit mode.

**Latency**

Latency was monitored through testing; Avg. Disk/sec reads and writes were monitored, at the following distances:

- 80 km synchronous-commit mode
- 800 km asynchronous-commit mode
- 4,000 km asynchronous-commit mode

After FAST Cache had warmed up, it allowed latency to be maintained at less than 3 ms for reads and 1 ms for writes on the primary replica for all three distance options, as shown in Table 11 (this table relates directly to Table 10, SQL Server LUN Design.)

### Table 11. Latency for reads and writes

<table>
<thead>
<tr>
<th>LUN ID</th>
<th>Name</th>
<th>Avg. Disk/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Reads</td>
</tr>
<tr>
<td>100</td>
<td>LUN_100_OLTP1_DATA_1</td>
<td>3</td>
</tr>
<tr>
<td>101</td>
<td>LUN_101_OLTP1_DATA_2</td>
<td>3</td>
</tr>
<tr>
<td>102</td>
<td>LUN_102_OLTP1_DATA_3</td>
<td>3</td>
</tr>
<tr>
<td>200</td>
<td>LUN_200_OLTP1_Log</td>
<td>0</td>
</tr>
<tr>
<td>103</td>
<td>LUN_103_OLTP2_DATA_1</td>
<td>3</td>
</tr>
<tr>
<td>104</td>
<td>LUN_104_OLTP2_DATA_2</td>
<td>3</td>
</tr>
<tr>
<td>105</td>
<td>LUN_105_OLTP2_DATA_3</td>
<td>3</td>
</tr>
<tr>
<td>201</td>
<td>LUN_201_OLTP2_Log</td>
<td>0</td>
</tr>
<tr>
<td>106</td>
<td>LUN_106_OLTP3_DATA_1</td>
<td>2</td>
</tr>
<tr>
<td>107</td>
<td>LUN_107_OLTP3_DATA_2</td>
<td>2</td>
</tr>
<tr>
<td>202</td>
<td>LUN_202_OLTP3_Log</td>
<td>0</td>
</tr>
<tr>
<td>108</td>
<td>LUN_108_OLTP4_DATA_1</td>
<td>1</td>
</tr>
<tr>
<td>203</td>
<td>LUN_203_OLTP4_Log</td>
<td>0</td>
</tr>
</tbody>
</table>
Latency on the secondary replicas was considerably higher, far above the 20 ms rule generally considered for data files and the 5 ms for logs. Consideration here should be given to the fact that no read activity was generated against an active secondary copy that would further increase latencies. This means careful consideration should be given to storage design for secondary replicas, because poorly allocated storage will impact not only the level of read activity that can be sustained, but also the ability of the secondary replicas to perform in the event of a failover and their transition to primary replica roles.

**Physical disk utilization**

Physical disk utilization was measured after analyzing the Unisphere NAR files, looking at physical disk utilization percentage, as shown in Figure 16.

![Physical disk utilization for storage pool on production and DR arrays](Image)

**Figure 16.** Physical disk utilization for storage pool on production and DR arrays

Initial disk utilization on the storage pool hosting the primary replica at the production site was too high at 90 percent. Improvements were seen after EMC FAST Cache was enabled. FAST Cache was able to reduce pressure on the SAS pool because frequently accessed data from the pool was placed in cache. After a two-hour warm-up period, disk utilization on production reduced from 90 percent to just 57 percent.

While FAST Cache boosts storage performance for SQL Server 2012, allowing the primary replica to service increased I/O levels, pressure on the storage pool hosting the data files for the secondary replica also increases as the writes on the primary are replicated to the secondary. This highlights the importance of correctly sizing the secondary replicas’ storage correctly. This solution has 40 SAS disks in both storage pools on the production and DR storage arrays. The differentiator was having FAST Cache enabled on the storage pool at the production site.
Storage processor utilization

Storage processor utilization was measured after analyzing the Unisphere NAR files, as shown in Figure 17.

Figure 17. SP utilization for SPA and SPB on production and DR storage arrays

The results for the production storage array show how SPA and SPB storage processor utilization increases as the array works to automatically boost performance through EMC FAST Cache technology. The SPs are analyzing and promoting the frequently accessed data. SP utilization for the DR storage array increases slightly as disk utilization rises due to the increase in write data being replicated from primary to secondary databases.

Creation of the availability groups can be done through scripting or in SQL Server Management Studio (SSMS) by either completing details for a New Availability Group or following the New Availability Group Wizard. The wizard also has an option to generate the script during the steps.

To test creation times we generated a script file to create the availability groups which required running through sqlcmd, the SQL Server command utility, using the –S server and –i file options.

sqlcmd –S [Servername] –i [filename]

Figure 18 shows a simplified process flow for creation of an availability group. During testing it was found that if adding multiple databases, it was best to back up, restore, and add secondary replicas to the availability groups one at a time before looping back to add additional databases. This avoids too long a gap between the restore process and databases being added to the availability group and synchronization process starting. If too long a gap occurred the process could fail.
The following data tables outline database creation times at:

- .8 ms (80 km) synchronous-commit
- 8 ms (800 km) asynchronous-commit mode
- 40 ms (4,000 km) asynchronous-commit mode

**Note:** These timings are for databases already populated and running a full workload of approximately 50,000 IOPS.

Availability group creation involved a COPY_ONLY full backup being taken, along with a transaction log backup, for each database included in the availability group, to a volume hosted on the production array. A restore occurs across the WAN to create the secondary replicas. The secondary databases remain in a synchronizing state unless configured for synchronous-commit mode, when the state changes to synchronized a short time after the availability group creation finishes.

**Table 12. Synchronous and asynchronous availability group creation at 80, 800, 4,000 km**

<table>
<thead>
<tr>
<th>Availability Group</th>
<th>Mode</th>
<th>Distance</th>
<th>Duration</th>
<th>Max Bandwidth (MB/sec)</th>
<th>Backup</th>
<th>Restore</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP_AG1</td>
<td>Synchronous</td>
<td>80 km</td>
<td>4hrs 14m</td>
<td>101.718</td>
<td>242.344</td>
<td></td>
</tr>
<tr>
<td>OLTP_AG2</td>
<td>Synchronous</td>
<td>80 km</td>
<td>4hrs 16m</td>
<td>109.683</td>
<td>254.839</td>
<td></td>
</tr>
<tr>
<td>OLTP_AG1</td>
<td>Asynchronous</td>
<td>800 km</td>
<td>4hrs 48m</td>
<td>101.63</td>
<td>179.683</td>
<td></td>
</tr>
<tr>
<td>OLTP_AG2</td>
<td>Asynchronous</td>
<td>800 km</td>
<td>8hrs 13m</td>
<td>110.78</td>
<td>189.489</td>
<td></td>
</tr>
<tr>
<td>OLTP_AG1</td>
<td>Asynchronous</td>
<td>4000 km</td>
<td>5hrs 50m</td>
<td>114.661</td>
<td>119.243</td>
<td></td>
</tr>
<tr>
<td>OLTP_AG2</td>
<td>Asynchronous</td>
<td>4000 km</td>
<td>8hrs 26m</td>
<td>100.541</td>
<td>259.641</td>
<td></td>
</tr>
</tbody>
</table>
The timings demonstrate that, as expected, as distance increases the time taken for creation of the availability groups increases. Availability group OLTP_AG1 contains only one 1 TB database, availability group OLTP_AG2 contained three databases (500 GB, 250 GB, 50 GB). The figures for bandwidth in MB/sec used were taken from the `sqlcmd` script window during creation. It should be noted that the configuration of the share space used for the Full Initial Data Synchronization can impact this process.

**Failover**

This solution tested the functionality of both manual and automatic failover on an AlwaysOn availability group by using SQL Server Management Studio. There are two types of manual failover, planned manual failover or forced manual failover (forced failover).

An availability group fails over at the level of an availability replica. If you fail over to a secondary replica in the SYNCHRONIZED state, the wizard performs a planned manual failover without data loss. If you fail over to a secondary replica in the SYNCHRONIZING, UNSYNCHRONIZED or NOT SYNCHRONIZING state, the wizard performs a forced manual failover—also known as a forced failover—with possible data loss. Both forms of manual failover transition the secondary replica to which you are connected to the primary role.

![Failover initiated through SQL Server Management Studio](image.png)

**Figure 19.** Failover initiated through SQL Server Management Studio

Forced manual failover can be initiated using the Failover Availability Group Wizard in SSMS, as shown in Figure 19. Alternatively, failover can be initiated through script.

After the wizard successfully completes the forced failover, the former primary replica comes online and transitions to the secondary role. The former secondary replica on the DR site in turn transitions to the new primary replica.

Automatic failover was also tested successfully as part of this solution. Automatic failover is available only between primary and secondary replicas that are configured for synchronous-commit mode and automatic failover mode when the secondary replica is in the SYNCHRONIZED state.
Conclusion

Summary

SQL Server 2012 AlwaysOn Availability Groups provides the flexibility to protect specific databases, individually or collectively, in either synchronous or asynchronous availability modes. These configurations allow SQL Server 2012 the ability to replicate data to a secondary replica over distance.

The solution clearly shows the ability of EMC FAST Cache to significantly boost performance of the VNX series storage array. Testing shows how enabling FAST Cache on a heavily utilized storage pool not only alleviated pressure, but allowed the same storage pool to service over **four times the I/O** (from 11,500 IOPS to 50,000 IOPS), while returning incredibly low SQL Server datafile average latencies times of **3 ms and less for reads and writes**. The ability of FAST Cache to automatically react to the changes in OLTP workload I/O patterns is an invaluable tool for administrators.

The solution also highlights the power of EMC VNX series to host active SQL Server 2012 environments, with databases protected by the new AlwaysOn Availability Group feature, in both synchronous and asynchronous modes, with minimal impact to performance.

While the performance improvements observed with these scenarios may not be representative of all Microsoft SQL Server 2012 environments, they nevertheless illustrate the potential to service far greater IOPS and reduce latency, using automated processes.

FAST Cache can also provide significant TCO savings by requiring far less drives than would be needed if FAST Cache was not enabled on the array. For example, by reducing power consumption and foot print, while at the same time simplifying the design and configuration for administrators required to optimize busy SQL Server 2012 environments.

Findings

Testing and validation demonstrated that:

- **SQL Server AlwaysOn Availability Group technology** brings significant improvements to the levels of native protection in terms of ease of use, monitoring, and performance compared to legacy native SQL Server protection offerings in previous versions.

- **EMC’s storage array platforms**, such as the EMC VNX, have been fully validated by EMC to support critical, enterprise-level, SQL Server 2012 deployments in a multi-subnet disaster recovery configuration.

- The VNX5700 can easily service **50,000+ SQL Server OLTP-like IOPS** while providing high availability through the SQL Server 2012 AlwaysOn Availability Groups configuration.
• EMC FAST Cache technology can be easily enabled to receive significant boosts in SQL Server 2012 transactional performance

Note The EMC VNX5700 array can handle far more IOPS than recorded in this white paper. The performance achieved in this solution was reached with the given amount of storage devices (refer to the Solution configuration section). Far higher levels of performance can be achieved in larger configurations.

Results clearly show the benefits of using EMC Technology with SQL Server 2012.
References

White papers
For additional information, see the white papers listed below:

*Microsoft SQL Server AlwaysOn Solutions Guide for High Availability and Disaster Recovery*

Other documentation
For additional information about Microsoft SQL Server 2012 and AlwaysOn, refer to the documents listed below:

- **Books Online for SQL Server 2012**
  - AlwaysOn Availability Groups
  - Availability Group Listener
  - AlwaysOn Dashboard
- **Pre-Configuration Database Optimizations**
- **SQL Server Best Practices article**