

MICROSOFT SQL SERVER BEST PRACTICES AND DESIGN GUIDELINES FOR EMC STORAGE

EMC VNX Family, EMC Symmetrix VMAX Systems, and EMC Xtrem Server Products

- Design and sizing best practices
- SQL Server performance acceleration with flash technologies
- Disaster recovery and high availability best practices

EMC Solutions

Abstract

This white paper identifies best practices and key decision points for planning and deploying Microsoft SQL Server with the EMC® VNX® family of unified storage, EMC Symmetrix® VMAX® series storage, and EMC XtremSF™ and EMC XtremSW™ Cache products.

October 2013



Copyright © 2013 EMC Corporation. All Rights Reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided “as is.” EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com.

All trademarks used herein are the property of their respective owners.

Part Number H12341

Table of contents

Executive summary	7
Purpose of this paper.....	7
Audience	7
Scope	7
Terminology.....	8
Microsoft SQL Server components and architecture	10
SQL Server overview.....	10
SQL Server releases.....	10
SQL Server 2012	10
SQL Server 2012 editions.....	10
SQL Server components.....	11
Windows Server	11
Windows volume types.....	12
SMB 3.0	13
SQL Server architecture.....	13
SQL Server logical components.....	15
SQL Server physical components.....	16
File types.....	16
Page and extent	16
Transaction log.....	17
Filegroup.....	17
I/O and bandwidth characteristics of SQL Server	18
Overview.....	18
OLTP	18
Data warehouse/OLAP database.....	19
Reading pages	20
Writing pages.....	20
Log Manager	21
Tempdb usage	21
I/O patterns	21
Best practices for SQL Server storage sizing and provisioning	23
Overview.....	23
General SQL Server storage best practices	23
Basic best practices for SQL Server.....	24
Basic best practices for storage.....	25
Clustering considerations.....	25
Considerations for earlier versions	26

General storage considerations.....	26
Performance versus capacity considerations	26
Disk type selection	26
Pools and RAID types.....	28
Virtual Provisioning storage considerations	30
Thin LUN versus thick LUN	30
Storage sizing best practices	31
Consideration for OLTP database sizing	31
Best practices for FAST VP sizing	32
Consideration for OLAP database sizing.....	34
Hypervisor storage considerations	35
General virtualization guidelines	35
Best practices for the VMware vSphere environment	38
Microsoft Hyper-V.....	40
SQL Server clustering storage considerations.....	41
Symmetrix VMAX storage design guidelines.....	41
VMAX series hardware design considerations	42
Virtual Provisioning considerations and best practices.....	42
FAST VP considerations and best practices for a VMAX storage system.....	42
VNX storage design guidelines.....	43
FAST Cache considerations and best practices	43
FAST VP considerations and best practices.....	44
FAST Cache versus FAST VP	45
Server flash considerations.....	45
XtremSF overview	45
Design best practices for XtremSF	45
XtremSW Cache overview	46
Design best practices for XtremSW Cache.....	47
Design best practices for XtremSW Cache in a virtualized environment	47
Sizing consideration for XtremSF and XtremSW Cache.....	48
Automation with ESI.....	49
SQL Server protection.....	51
Overview.....	51
AlwaysOn Availability Groups.....	51
SQL Server native data protection	52
Recoverable versus restartable copies.....	52
VDI and VSS frameworks for backup replication	53
EMC high availability and data protection offerings for SQL Server.....	53
Replication technologies	55
Replication management tools	56
Multi-site disaster recovery	56

Considerations.....	56
Multi-site replication technologies	57
Tools to restart automation	58
Virtualized instances automation tools	58
Disaster recovery options for SQL Server 2012.....	59
Additional backup recommendations.....	59
AlwaysOn for HA/DR.....	59
AlwaysOn with FAST Suite	59
AlwaysOn with flash XtremSW Cache/XtremSF	60
Conclusion	61
Summary	61
Additional information	61
Appendix A: EMC Data Protection Advisor for Replication Analysis	62
Overview.....	62
Data Collection and Discovery wizards	62
Data discovery and collection	62
Discovering storage arrays	62
Configuring Data Protection Advisor for Microsoft SQL Server monitoring	63
Displaying and reporting gaps and exposures	64
Appendix B: Tools for SQL Server performance monitoring, tuning, and sizing	66
Overview.....	66
Application-level tools	68
EMC DBclassify.....	68
Perfcollect	70
PAL.....	70
SQL Server database-level tools.....	70
VSPEX SQL Server sizing tool.....	70
Transact-SQL.....	71
SQL Server Profiler.....	72
SQL Server Database Engine Tuning Advisor.....	72
SQL Server Dynamic Management Views	73
Windows host-level tool.....	73
Windows Performance monitor (Perfmon).....	73
Hypervisor-level tools	74
Key metrics to monitor ESX.....	74
Key metrics to monitor Hyper-V.....	76
Storage/Server cache-level tools.....	77
Unisphere Analyzer	77
XtremSW Cache Performance Predictor.....	77

EMC Storage Configuration Advisor	80
Appendix C: SQL Server Workload generation tools	83
Overview.....	83
Tools introduction.....	83
SQL Server Profiler.....	83
IOMeter	84
SQLIO.....	84
SQLIOSim.....	84
Quest Benchmark Factory.....	84
Appendix D: Sample storage designs and reference architectures	85
Overview.....	85
Microsoft SQL Server storage design on VMAX with FAST VP.....	85
Phase 1–Collect user requirements	85
Phase 2–Design the storage architecture based on user requirements	85
IOPS calculation	86
Capacity calculation	86
Building-block design approach for data warehouse.....	88
Building-block design considerations.....	88
Building-block design details	89
Deploying building blocks	91
SQL Server virtual machine and LUN allocation design	92
SQL Server protection solution.....	93
EMC RecoverPoint	93
EMC Replication Manager.....	102
VMware vCenter SRM	104

Executive summary

In the planning and design phases of a Microsoft SQL Server implementation, it is important to understand how the application interacts with the storage platform. It is also critical to know storage-design best practices to avoid problems and achieve high performance.

From a storage design perspective, consider the application architecture and user profile characteristics of Microsoft SQL Server for performance, protection, and growth of the SQL Server database.

This paper can help solution professionals assess and address SQL Server storage requirements for performance, scalability, and availability:

- It is always preferable to collect actual data from the site.
- In the absence of the actual performance data, make a series of reasonable assumptions when designing for a typical environment.
- Always consider protection requirements when designing a storage system.

Purpose of this paper

This paper presents the set of current EMC-recommended best practices for storage design in support of Microsoft SQL Server. Guidelines are presented within the context of deploying SQL Server on the EMC® VNX® family, EMC Symmetrix® VMAX® series, and EMC Xtrem™ family. The paper includes guidelines for deploying SQL Server in both physical and virtual environments.

Audience

This white paper is intended for customers, EMC partners, and service personnel who are considering implementing a database environment with Microsoft SQL Server or considering upgrading an earlier version of SQL Server. We assume that the audience is familiar with Microsoft SQL Server, EMC storage products such as VNX, Symmetrix VMAX, XtremSF™, and XtremSW™ Cache, as well as VMware or Microsoft Hyper-V virtual environments.

Scope

This document presents storage design best practices recommended by EMC for hosting Microsoft SQL Server on EMC VNX storage, EMC Symmetrix VMAX storage, and XtremSF or XtremSW Cache in both physical and virtual environments. The paper includes sizing and design examples based on EMC's proven approaches. Detailed, end-to-end implementation instructions are beyond the scope of this document.

Terminology

This white paper includes the following terminology.

Table 1. Terminology

Term	Definition
Availability Groups (AG)	A high availability (HA) and disaster recovery feature in SQL Server 2012. Maximizing the availability of a set of user databases, it provides an enterprise-level alternative to database mirroring.
Availability replica	An instance 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 (see <i>Primary replica</i> in this table) and up to four secondary replicas (see <i>Readable secondary replica</i>).
Data synchronization	The process by which changes to a primary database are reproduced on a secondary database.
EMC XtremSF	A single, low-profile server flash hardware card that fits in any rack-mounted server within the power envelope of a single PCIe slot, available with a broad set of enterprise multi-level cells (eMLC) and single-level cell (SLC) capacities.
eMLC	Enterprise multi-level cell. A multi-level cell is a flash memory technology designed for low error rates using multiple levels per cell to allow more bits to be stored using the same number of transistors.
FAST™ Cache	Fully Automated Storage Tiering (FAST) Cache is EMC software that enables customers to add various flash drive capacities in order to extend existing cache capacity for better system-wide performance. FAST Cache is now available with increased capacity configurations using the 100 GB flash drive or the 200 GB flash drive. These additional configurations are available only on the VNX storage array.
Fully Automated Storage Tiering for Virtual Pools (FAST VP)	A feature of VNX storage arrays that automates the identification of data volumes for the purpose of allocating or reallocating business application data across different performance and capacity tiers within the storage array.
Multi-level cell (MLC) flash	A flash memory technology using multiple levels per cell to allow more bits to be stored using the same number of transistors.
NAND	NAND flash memory is a type of non-volatile storage technology that does not require power to retain data.
OLTP	Online transaction processing. Typical applications of OLTP include data entry and retrieval transaction processing.

Term	Definition
Primary replica	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.
RAID	Redundant array of independent disks (RAID) is a method of storing data on multiple disk drives to increase performance and storage capacity and to provide redundancy and fault tolerance.
Readable secondary replica	Secondary replica databases configured to allow read-only client connections.
Reseeding	Process of copying a database from a primary replica to corresponding secondary replicas.
Single-level cell (SLC) flash	A type of solid-state storage (SSD) that stores one bit of information per cell of flash media.
SP	Storage processor.
SQL Server 2012 AlwaysOn	A comprehensive high availability 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.
Storage pool	Virtual constructs that enable data to move dynamically across different tiers of storage according to the data's business activity. With VNX and VMAX systems, storage pools are fully automated and self-managing.
Thin LUN	A type of LUN created in storage pool in which the physical space allocated can be less than the user capacity seen by the host server.
Thick LUN	A type of LUN created in storage pool in which the physical space allocated is equal to the user capacity seen by the host server
VMDK	Virtual Machine Disk file format in an ESXi Server.
VHDX	Virtual Hard Disk format in Windows Server 2012 Hyper-V.

Microsoft SQL Server components and architecture

SQL Server overview

Microsoft SQL Server is Microsoft's relational database management and analysis system for day-to-day operation and data-warehousing solutions. The current version is Microsoft SQL Server 2012, and previous versions include Microsoft SQL Server 2008 R2, SQL Server 2008, SQL Server 2005, and SQL Server 2000.

SQL Server releases

In the release of **SQL Server 2000**, Microsoft focused on the development of Business Intelligence functionality including the extract, transform, and load (ETL) tool, Reporting Server, and online analytical processing (OLAP) analysis Services.

SQL Server 2005 introduced the XML data type, Dynamic Management Views (DMVS) for monitoring and diagnosing server state and performance, as well as Common Language Runtime (CLR) to integrate with .NET Framework. SQL Server 2005 Service Pack 1 (SP1) added Database Mirroring for redundancy and failover capability at the database level.

SQL Server 2008 introduced AlwaysOn technologies to reduce downtime, aiming to make data management self-tuning, self-organizing, and self-maintaining. SQL Server 2008 R2 added Master Data Services to centrally manage master data entities and hierarchies and Multi-Server Management to centralize multiple SQL Server instances and services.

SQL Server 2012 introduced AlwaysOn SQL Server Failover Cluster instances and Availability Groups to improve database availability, columnstore indexes to increase query performance, Contained Databases to simplify the moving between database instances, and better memory management.

Each version of SQL Server comes in various editions, which can be considered as a subset of the product features. Users can verify which edition they are running with the query: `select serverproperty('edition')`. The mainstream editions include Datacenter edition, Enterprise edition, Standard edition, Web edition, Business Intelligence edition, Workgroup edition, and Express edition.

SQL Server 2012

SQL Server 2012 is the latest version of Microsoft SQL Server. It supports high availability and disaster recovery through AlwaysOn clusters and availability groups, xVelocity in-memory storage for fast query performance, rapid data exploration through PowerView and tabular modeling in Analysis Services, and new data management capacity with Data Quality Services.

SQL Server 2012 editions

Microsoft SQL Server 2012 includes the following principal editions:

- **SQL Server Standard Edition:** This edition delivers basic data management and Business Intelligence reporting and analytics capacities. It provides effective database management with minimal IT resources.

- **Business Intelligence Edition:** In addition to all the capabilities of SQL Server standard edition, this edition also supports self-service and scalable BI solutions. It features:
 - **PowerView:** An add-in feature for SQL Reporting services for rapid data discovery
 - **PowerPivot:** A feature that easily collaborates and shares insight with access and mash up data
 - **Master Data Services:** Used for maintaining master data used for object mapping, reference data, and metadata management across the organization structure
 - **BI semantic model:** Provides a consistent view across heterogeneous data sources and transforms user-created applications into corporate BI solutions
- **Enterprise Edition:** This edition delivers comprehensive high-end data center capability. It can handle demanding workloads with fast performance while still maintaining required uptime and data protection. It features:
 - **SQL Server AlwaysOn:** Provides greater uptime, faster failover, and better use of hardware resources with unified high availability solution
 - **PowerView:** Creates and interacts with views of data from data models based on PowerPivot workbooks and provides intuitive ad-hoc reporting
 - **xVelocity:** Uses columnstore storage with memory caching, highly parallel data scanning, and aggregation algorithms to boost performance across data warehousing and Business Intelligence
 - **Data Quality Services:** Improves data quality by using organizational knowledge and third-party reference data providers to profile, cleanse, and match data

SQL Server components

SQL Server consists of four key components:

- **SQL Server Database Engine:** Creates and drives relational databases
- **SQL Server Integration Services (SSIS):** Performs Extract, Transform, and Load (ETL) process to clean up and format raw data from source systems into the databases as ready-to-use information
- **SQL Server Analysis Services (SSAS):** The data analysis component that creates OLAP cubes and data mining
- **SQL Server Reporting Services (SSRS):** Provides a reporting framework to create, manage, and deploy tabular matrix graphical reports

Windows Server

The database platform is closely related to the operating system. Microsoft Windows Server provides a solid infrastructure for SQL Server.

Windows volume types

Windows volume partition styles include MBR and GPT:

- **MBR:** The legacy partitioning style, which allows a maximum of four partitions. The partition table is saved only at the beginning of the disk.
- **GPT:** A partitioning style capable of managing partitions larger than 2 TB. Its partition table is saved in multiple locations. It can be easily recovered if any partition is corrupted.

Two types of disk modes are supported:

- **Basic:** The most basic disk, which contains primary partitions and if needed, extended partitions. Basic mode's features include:
 - Primary partition: A standard bootable partition.
 - Extended partition: A non-bootable partition. This is the fourth partition on a Basic MBR disk, which holds logical partitions, thus allowing more than four partitions.
 - Logical partition: A non-bootable partition contained in the extended partition to extend the basic disk.
 - Extensible firmware Interface (EFI): Used to store boot files on EFI compatible systems.
 - Microsoft System Reserved (MSR): Only available on GPT basic disks, used to reserve space for future use.
- **Dynamic:** Dynamic disk is a native host-based logical volume manager, responsible for aggregating disks into logical volumes with multiple options. It creates two partitions; one contains all dynamic volumes and the other is hidden and contains the Logical Disk Manager (LDM) database. This database is replicated across all dynamic disks on the system so it is recoverable. It can host up to 2,000 dynamic volumes (a maximum of 32 is recommended). Dynamic mode's features include:
 - Simple: Standalone volume
 - Striped: Like RAID 0, striped volume writes a data block to both disks. The volumes integrating this arrangement must be the same size.
 - Spanned: Like RAID 0, with concatenated volumes. If the disk fails, only part of data will be lost. The volumes are not required to be of the same size. It has lower performance than striped volumes with the same amount of disks.
 - Mirror: RAID 1
 - RAID: RAID 5

Table 2 describes the typical volumes created on EMC storage and used in the SQL Server environment.

Table 2. Typical SQL Server deployment for EMC storage

Volume partition	Disk	Volume	Allocation size	Formatting options
MBR	Basic	NTFS	64 KB	Quick format*

Note: Because the EMC array provides storage RAID protection, Dynamic disks should be avoided if possible as it complicates the management of storage as well as local and remote disaster recovery (DR). Quick Format options are required for thin LUNs.

SMB 3.0

Server Message Block (SMB) 3.0 is a new version of the existing network file sharing protocol that allows applications on a computer to read and write to files and to request services from server programs in a computer network.

SMB 3.0 was introduced in Windows Server 2012 and has been supported by SQL Server 2012 as a viable storage topology for Databases since SQL Server 2012's RTM release.

SQL Server 2012 supports both virtualized disk (VHD/VHDX) and databases directly hosted on SMB 3.0 shares. The shares can be presented to Windows Server 2012 or multiple cluster servers.

SMB 3.0 provides the ability to survive hardware failures that otherwise impact file access. EMC provides full support for SMB3.0 as an NFS storage topology for SQL Server.

Refer to [Storage Windows 2012](#) for detailed windows storage descriptions.

SQL Server architecture

Figure 1 shows the four major components of the SQL Server architecture: SQL OS, storage engine, query processor, and protocol layer.

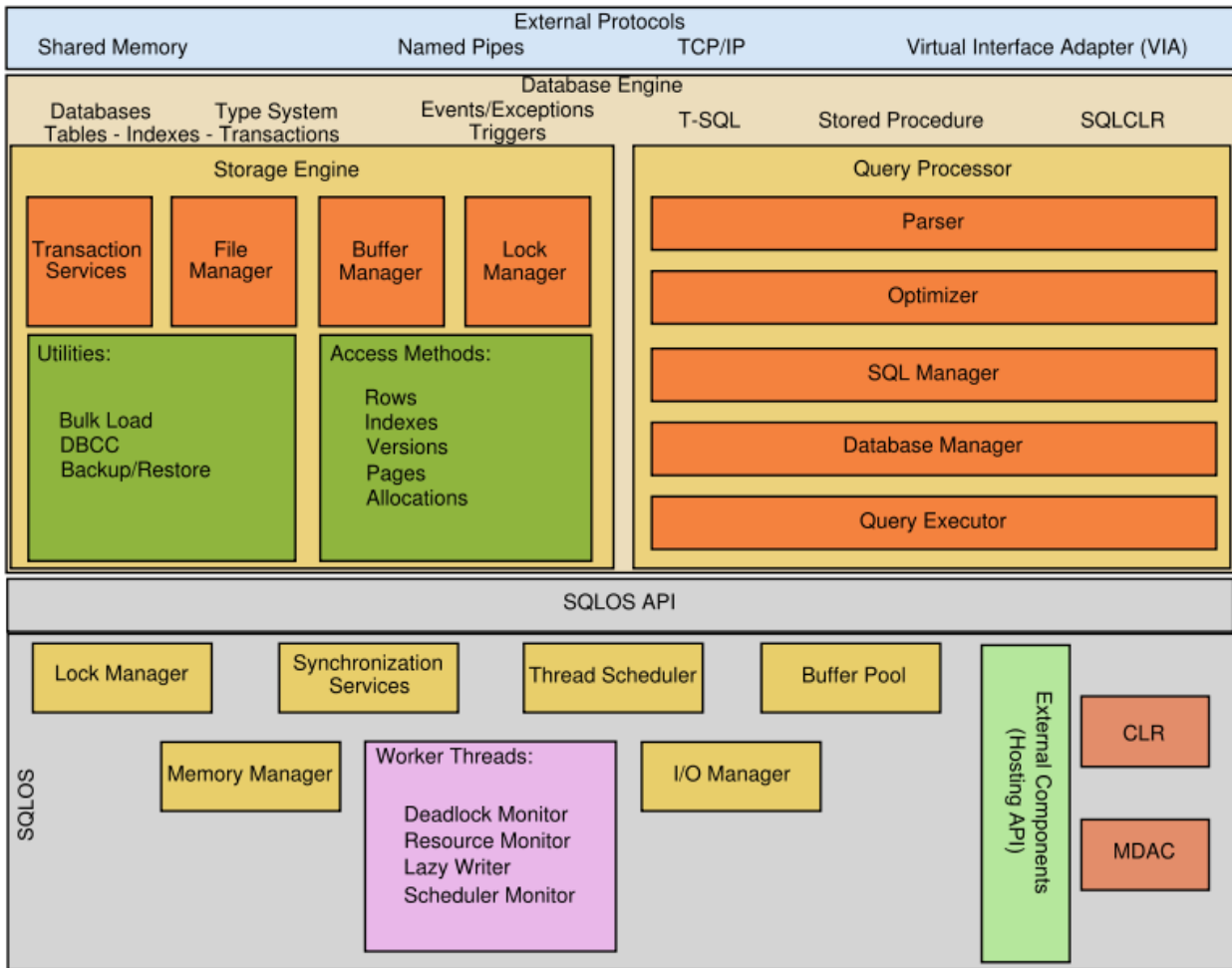


Figure 1. SQL Server architecture

- **SQL OS** is the application layer at the lowest level of the SQL Server database engine. It handles activities such as scheduling, deadlock detection, and memory management.

SQL Server manages its memory resources dynamically. The buffer pool is the main memory component in SQL Server. Memory that is not used by other memory components remains in the buffer pool and will be used as data cache for pages read from the database files on the disk. The memory manager manages disk I/O functions for bringing the data and index pages into the data cache so that the data can be shared among users.

- The **storage engine** manages all data access through transaction commands and bulk operations. It has three main areas: access methods, locking and transaction service, and utility commands.

- The **query processor (relational engine)** layer accepts T-SQL batches and determines what to do. It parses, compiles, and optimizes the T-SQL queries requests and oversees the process of executing the batch. As the batch is executed, a request for the data is passed to the storage engine. The query processor has two components: Query Optimizer and Query Executor.
 - **Query Optimizer** determines the best execution plan.
 - **Query Executor** executes the query.

The query processor also manages the execution of queries that request data from the storage engine and processes the returned results. Object Linking and Embedding Database (OLE DB) row set is the communication channel between the relational engine and the storage engine.

The command parser handles T-SQL language events sent to the SQL Server instances, checks for proper syntax, and translates T-SQL commands into the query tree. The Query Optimizer takes the query and prepares it for execution by compiling the command batch, optimizing the queries, and finding the best way to process it in an execution plan. The Query Executor runs the execution plan, acting as a dispatcher for all the commands in the execution plan.

- The **protocol** layer receives the request from the user application and translates it into a form that the relational engine can work with. It also translates the results of queries, status, and error messages into a format the client can understand.

SQL Server logical components

Microsoft SQL Server includes two main logical components:

- **Relational engine** (query processor), used to verify SQL statements and select the most efficient way to retrieve the query data
- **Storage engine**, used to execute physical I/O request and return the row requested by relational engine

These two engines work together to provide data integrity for SQL Server.

The SQL Server logical architecture defines how the data is logically grouped and presented to the users. The following are the core components in this architecture:

- **Tables:** Tables are formed with logically aggregated data pages (the basic format for data). Columns and rows are the two main components in a SQL Server table.
- **Indexes:** An index created on one or more columns of a table and associated with a table or view speeds up data retrieval. Clustered and non-clustered indexes are supported. A table can have only one clustered index that defines the order in which the data is stored in the table. A heap table is a table with no index.
- **Views:** A view can be a virtual table or a stored query. The data returned from a view is stored in the database through the selected statement.
- **Stored procedure:** A stored procedure is a group of Transact-SQL statements compiled into a single execution plan.

- **Constraints, Rules, and Triggers:** These are components that are used to maintain the data type and the data integrity of the table.
- **User-defined functions:** Functions are used to encapsulate frequently performed logic.
- **Triggers:** A trigger is similar to a stored procedure. It is attached to a table and executed only when an INSERT, UPDATE, or DELETE command triggers it.

SQL Server physical components

The SQL Server physical components determine how the data is stored in the file system of the operating system. Database files, page, extent, and transaction log files are core physical components of SQL Server.

File types

SQL Server databases have the following file types:

- **Primary data files** have an MDF extension. A database requires at least one primary data file.
- **Secondary data files** have an NDF extension. All data files in a database that are not primary data files are secondary data files. Secondary data files are not required, and a database can have many secondary files or none.
- **Log files** have an LDF extension. They hold all the transaction log information needed to recover the database. Each database has one log file regardless of the number of data files.

Data files store data and index information. Figure 2 represents the physical layout of a single data file object showing the relationship of pages and extents.

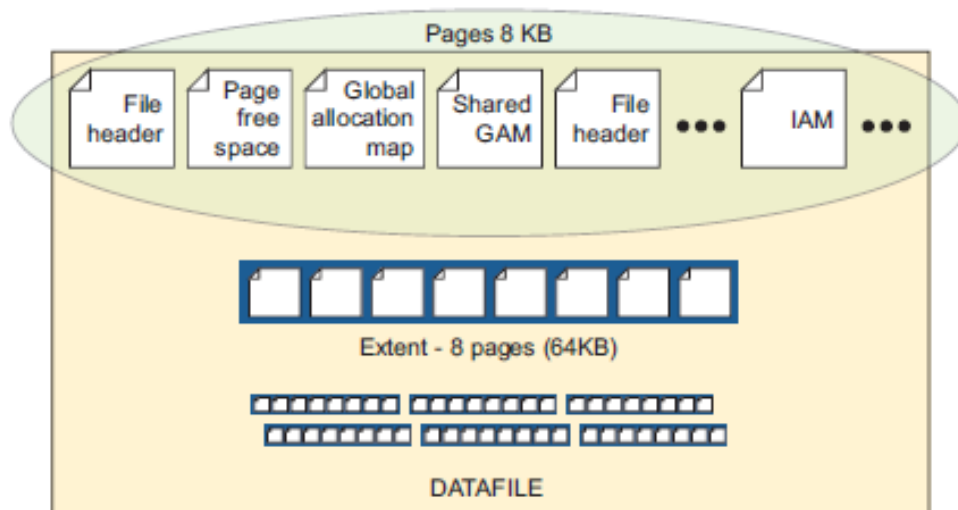


Figure 2. Data file, pages, and extents

Page and extent

A SQL Server **page** is the basic unit of logical data storage. With a page size of 8 KB (128 pages per megabyte), each page begins with a 96-byte header containing system information about the page.

The disk space allocated to the primary or secondary data file (.mdf or .ndf) is logically divided into pages. Disk I/O operations are performed at the page level.

Extents are the basic units that manage the space. Each extent has eight physically adjacent pages, which is 64 KB (16 extents per megabyte). A table or index is usually allocated with pages from mixed extents. Uniform extents are used for subsequent allocations after growing to eight pages.

Transaction log

The transaction log maintains modifications made by transactions within the data files. It contains information regarding the following events:

- The start and end of each transaction
- Data modification
- Extent and page allocation and de-allocation
- Creation and elimination of a table or index

The transaction log is critical to recover databases during a system failure.

Log records are stored in a serial sequence and each record contains a transaction ID. A single physical transaction log is logically segmented into virtual logs based on internal SQL Server algorithms and the initial size of the transaction log. A virtual log within the physical log file records transactional information whenever transactional activity begins.

Filegroup

A SQL Server filegroup can be used to separate files for tables and indexes, allowing selective placement of them at the disk level. It can

- Separate tables and indexes at the disk level
- Separate the objects that require more data files because of a high page allocation rate

The SQL Server database administrator can:

- Perform a backup at the filegroup or file level. SQL Server is able to provide partial availability to a specific filegroup. It can be online as long as the primary filegroup is online, even when other filegroups are offline. A filegroup is available if all its files are available.
- Use separate filegroups for in-row data and large-object data in tables and indexes
- Use filegroup for partitioned tables:
 - Each partition can be in its own filegroup
 - Partitions can be switched in and out of the table for better archiving

I/O and bandwidth characteristics of SQL Server

Overview

Understanding the SQL Server I/O pattern and characteristics is critical for designing and deploying SQL Server applications. A properly configured I/O subsystem can optimize SQL Server performance.

There are two types of generic SQL Server database workloads: OLTP and data warehouse/OLAP. A specific user database might generate an I/O workload that is drastically different from those in the standard benchmark. The only way to determine I/O performance needs is to analyze the database under a typical load in real-time.

OLTP

OLTP workloads produce numerous concurrent transactions with significant random I/O reads and writes (IOPS). OLTP databases change constantly. Most ad-hoc applications generate OLTP workload.

According to [Microsoft SQL Server Best Practice](#) articles, OLTP database workloads contain the following patterns:

- Both reads and writes issued against data files are generally random in nature.
- Read activity (in most cases) is constant in nature.
- Write activity to the data files occurs during checkpoint operations (frequency determined by recovery interval settings).
- Log writes are sequential in nature with a varying size depending on the nature of the workload (sector aligned up to 60 KB).
- Log reads are sequential in nature (sector aligned up to 120 KB).

OLTP databases usually have many write activities that place pressure on the I/O subsystem, particularly on the log LUN, because the write goes to the transaction log first.

A typical OLTP system has a large number of concurrent connections actively adding and modifying data, for example, in an online airline reservation system. An OLTP system requires frequently backing up transaction logs and places further demands on the I/O subsystem.

In configurations that use the transactional replication, after the initial snapshot takes place, subsequent data changes and schema modifications made at the publisher are delivered to the subscriber, driving more read activity for the transaction log on the publisher database.

Index usage is another factor that affects the I/O subsystem. Heavily indexed OLTP systems can support high concurrency with low latency to retrieve a small number of rows from datasets that contain very little historical data. The volatility of transactions within an OLTP system might require frequent index maintenance that places heavy read and write requests on the I/O subsystem.

OLTP systems usually generate a large number of input/output operations per second (IOPS). More disk drives support more IOPS capacity.

Data warehouse/OLAP database

Data warehousing is often the basis of a decision support system (DSS) or a Business Intelligence system. It is a repository of an organization's data, designed to facilitate complex analytical query activities using very large data sets for reporting and analysis. Data warehouse databases are Online Analytical Processing (OLAP)-type databases, which typically use complex analysis with aggregated or summarized data in the data warehouse.

Data in the data warehouse system is usually static with sequential read and very little write activity, except for typical batch updates. I/O bandwidth is more important than IOPS. The typical workload in a data warehouse is I/O intensive with operations such as large data load and index build, creation of views, and queries over large volumes of data. The underlying I/O subsystem for the data warehouse should meet these high bandwidth requirements.

The I/O characteristics for data warehouse are:

- Sequential reads and writes, generally the result of table or index scans and bulk insert operations
- Non-volatile data, larger historical datasets
- Light index in the Fact table
- Low concurrency
- High Tempdb activity
- Varied I/O size: typically larger than 8 KB. Read-ahead is any multiple of 8 KB up to 512 KB. Bulk load operations are any multiple of 8 KB up to 128 KB.
- When using columnstore indexing, database file I/O size that is much larger than 256 KB

A key design consideration for an efficient data warehouse (DW) storage solution is to balance the capabilities of the DW system across the compute, network, and storage layers.

For example, the compute layer should be capable of processing data at bandwidth rates the storage can provide at comfortable utilization levels. In turn, the networking of compute and storage layers should be sufficient to sustain the maximum permissible throughput between compute and storage layers. Ideally, to ensure a cost-efficient DW solution, one element of the solution should not have an excessive capability over the other.

When designing a data warehouse, estimate how much I/O bandwidth a given server and host bus adapter (HBA) cards can potentially utilize, and ensure that the selected I/O configuration will be able to satisfy the server requirement.

A well-designed data warehouse system optimizes the storage system for scanning-centric operations; the server CPU can receive and process the data delivered by the storage at the same bandwidth. Because the queries in the data warehouse can fetch millions of records from the database for processing, the data is usually too large to fit in memory. A good storage design should quickly locate and deliver the data from disk for the processors to perform the aggregation and summarization.

Reading pages

I/O reads from the SQL Server database engine are of the following types:

- **Logical read:** Occurs when the database engine requests a page from the buffer cache
- **Physical read:** Copies the page from the disk into the cache if the page is not currently in the buffer cache

The read requests are controlled by the relational engine and optimized by the storage engine. The read-ahead mechanism anticipates the data and index pages needed for a query execution plan and brings the pages into the buffer cache before being used by the query. This mechanism makes it possible to overlap computation with I/Os to make full use of both CPU and disk to optimize the performance.

Writing pages

I/O writes from an instance of the database engine are of following types:

- **Logical write:** Occurs when data is modified on a page in buffer cache
- **Physical write:** Occurs when the page is written from buffer cache to the disk

Both page reads and page writes occur at the buffer cache. Each time a page is modified in the buffer cache, it is marked as “dirty.” A page can have more than one logical write before it is physically written to disk. The log records must be written to disk before the associated dirty page is written to disk. To ensure data consistency, SQL Server uses write-ahead logging to prevent writing a dirty page before the associated log record is written to disk.

Figure 3 illustrates the page write operation in SQL Server.

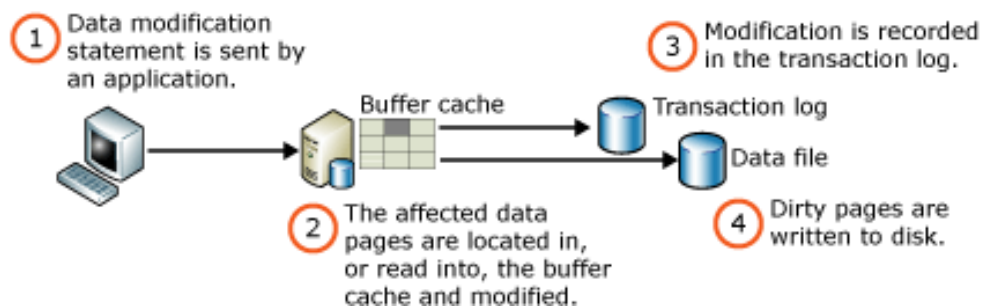


Figure 3. Page write operation in SQL Server

A dirty page is written to a disk in one of the following ways:

- **Lazy writing:** A system process that keeps free buffers available by removing infrequently used pages from the buffer cache. Lazy writing first writes dirty pages to the disk.
- **Eager writing:** This system process writes dirty pages with non-logged operations such as bulk insert or select.
- **Checkpoint:** The checkpoint operation periodically scans the buffer cache for database pages and writes all dirty pages to the disk.

The lazy writing, eager writing, and checkpoint processes use asynchronous I/O that allows the calling thread to continue processing while the I/O operation takes place in the background to maximize both CPU and I/O resources for the appropriate tasks.

Log Manager

The log workload is the I/O against the transaction log. It usually has sequential writes and requires low latency for high scale transaction workloads. Transaction log file writes are synchronous for a given transaction, because SQL Server flushes all updates associated with a committed transaction to the log before the user thread can start the next transaction.

Tempdb usage

Tempdb is a system database used by SQL Server as a temporary workspace. The I/O pattern for Tempdb is similar to OLTP patterns. Depending on the workload, Tempdb can range from low (in OLTP type of workloads) to extremely high activity (DSS or OLAP workloads).

I/O patterns

Table 3 summarizes the I/O patterns involved in each type of database.

Table 3. I/O patterns of different workloads for a SQL Server database

I/O types and characteristics	Database file Online Transaction Processing (OLTP)	Decision support system (data warehouse, OLAP)
Data files	<ul style="list-style-type: none"> • Smaller Random I/Os (8–64 KB) • High proportion of reads compared to writes (typically 90/10 to 70/30 read/write ratio) • High performance and protection can usually be achieved with RAID 10. With tiered storage, RAID 5 or RAID 6 can be used in the storage pool to provide sufficient performance. 	<ul style="list-style-type: none"> • Larger sequential I/Os (mostly 64 KB, can be more than 256 KB with columnstore index) • Low proportion of writes compared to reads, sometimes read-only • RAID 5 usually provides adequate performance and much more usable space for a given number of disks
Database log file	<ul style="list-style-type: none"> • Small, highly sequential I/Os (some multiples of 512 bytes) • Almost exclusively writes, with occasional reads during large rollbacks or log backups • RAID 1/0 recommended for logs. RAID 5 may also provide adequate performance (because of full stripe writes). The performance may decrease when there is a drive failure (the performance downgrade can be ignored if on flash drives.) 	
Tempdb data file	<ul style="list-style-type: none"> • Varying size depending on usage (usually larger I/O, typically does not exceed 64 KB) • Serial or random I/Os, a given workload might be somewhat sequential, many workloads running simultaneously may give Tempdb a random I/O appearance • Usually near 50/50 split of writes and reads • Based on the unpredictable nature of Tempdb combined with its usually large proportion of writes, RAID 1/0 usually provides the best performance for a given number of disks. Similar to log files, RAID 5 may also provide adequate performance, especially when flash drives are used. 	

I/O types and characteristics	Database file Online Transaction Processing (OLTP)	Decision support system (data warehouse, OLAP)
	<ul style="list-style-type: none"> • Tempdb activity varies. • Usually is not very active with low performance demand. • Can be very active for frequent reporting and large table joins. 	Tempdb can have high performance demand that requires server-side flash storage, such as XtremSF.

Best practices for SQL Server storage sizing and provisioning

Overview

Storage design is one of the most important elements of a successful Microsoft SQL Server deployment. To achieve a storage design for optimal reliability, performance, cost, and ease of use, follow the recommended storage guidelines.

This section provides general best practices for deploying SQL Server on EMC storage, such as Symmetrix VMAX series storage, VNX unified storage, XtremSF, and XtremSW Cache and recommendations for specific EMC storage array features with SQL Server.

Because the virtualization of a SQL Server environment requires its own set of considerations, this section also includes guidance on this subject.

General SQL Server storage best practices

EMC recommends that you start a SQL Server design with five LUNs, as shown in Figure 4, and expand based on application performance requirements.

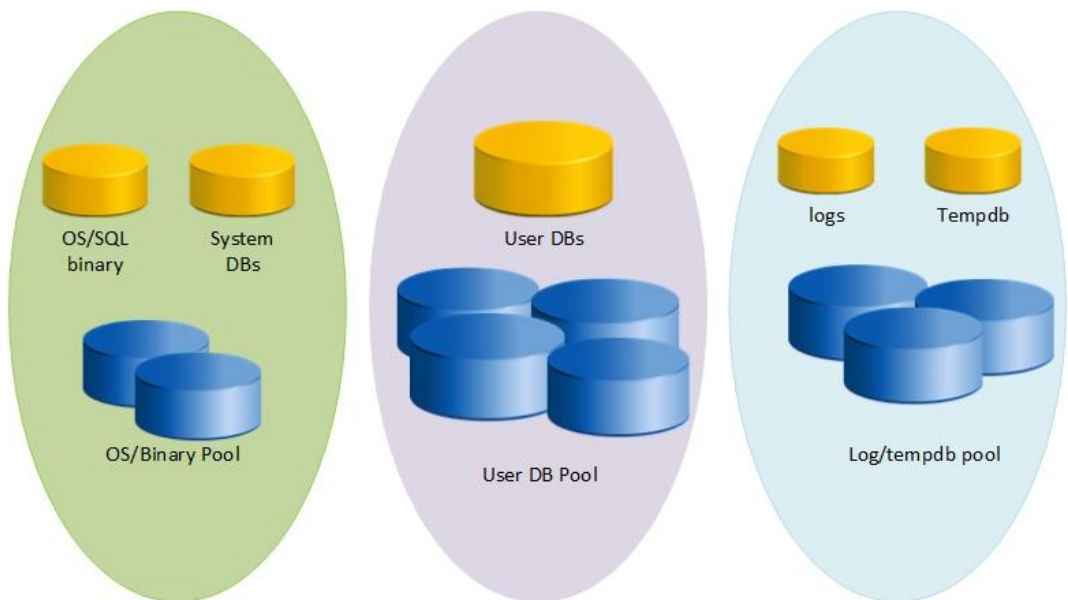


Figure 4. The SQL Server configuration

For the SQL Server configuration, begin by considering the following basic requirements:

- **OS/SQL Server binaries**

In a typical SQL Server implementation, the server is dedicated for SQL Server and binaries are on the same LUN where the OS is. Follow Microsoft's recommendation for the OS type and SQL Server version and consider the overhead for applications that you need to install on that server.

Typical LUNs for the OS/SQL Server binaries/system databases is 60-120 GB. Typically, high capacity/low performance disks in a RAID 5 storage pool can meet this need.

- **System databases**

In most environments, system databases are not frequently changed or modified and they can be on the same LUN as the OS.

- **Logs for user databases**

Logs for user databases typically need low IOPS (mostly sequential writes). Even with replication such as AlwaysOn Availability Group (AAG), the IOPS needed on these LUNs are typically not very demanding. Thus, the log LUNs are usually configured with Fibre Channel (FC) disks (in a storage pool, it can be pinned to the FC tier) that can satisfy the capacity need with at least 10 percent extra space.

- **Tempdb**

In an OLTP environment, Tempdb may not be very I/O demanding and can follow the same design principle for logs. In this case, they can usually be in the same pool with SQL Server database Log LUNs.

When there is scheduled or ad-hoc reporting or large table joins, Tempdb can potentially experience heavy usage. You must measure the SQL system needs to determine the Tempdb usage.

A Tempdb in a data warehouse or for OLAP workloads is typically under very intensive I/O demands and warrants special attention in these environments. The Tempdb design in these environments should follow the database design principle for sizing and placement if needed.

- **User database**

User database LUNs are typically the main focus for storage design. LUN types vary depending on performance and capacity requirements as well as the workload type.

Follow the general SQL Server storage best practices at [Microsoft TechNet](#). The following are some key points:

Basic best practices for SQL Server

The following are some basic SQL Server best practices:

- Select the **Lock pages in memory** policy for the SQL Server start account to prevent SQL Server from swapping memory.
- Pre-allocate data files to avoid **Autogrow** during peak time.
- Set **Autoshrink** to **Off** for data and log files.
- Make data files of equal size in the same database—SQL Server uses a proportional fill algorithm that favors allocations in files with more free space.
- Perform routine maintenance with index rebuild or reorganization with the command **dbcc checkdb**.

Filegroup and file considerations

The following are considerations regarding filegroup in SQL Server:

- Filegroups can be accessed in parallel; placing filegroups on different set of disks/storage pool can improve performance.

- Organize SQL Server data files with similar performance and protection needs into a filegroup when designing a database.
- For allocating intensive databases such as Tempdb, create 0.25 to 1 data files per filegroup for each CPU when needed.
- Start with a small number of data files. Increase the numbers when needed.
- Create one log file in a typical environment. More log files will not improve performance.

Refer to *Using Files and Filegroups* at the [Microsoft MSDN Library](#) for more information.

Basic best practices for storage

The following are some high-level basic best practices for storage design. The details are discussed in [General storage considerations](#).

- Plan for performance, capacity, and protection. Table 4 lists the response times for data file and log files.

Table 4. Response times for data file and log files

I/O response time	Data file	log
Very good	Less than 10 ms	Less than 5 ms
Acceptable	10 to 20 ms	5 to 15 ms
Needs investigation and improvement	Greater than 20 ms	Greater than 15 ms

- When creating a volume in windows, set the Windows allocation unit to 64 K for SQL Server database and log LUNs.
- For optimal performance with a predictable level of service, place Tempdb, data, and log files on separate LUNs.
- To leave room for data growth, avoid exceeding 80 percent capacity of the LUNs for database files.
- Place LUNs for data files on faster drives or use auto-tiering for them.
- Place LUNs for log files on SAS or FC drives without auto-tiering.
- Use up-to-date manufacturer-recommended HBA drivers.
- Ensure that storage array firmware is up-to-date.
- Consider using multipathing for availability/redundancy and throughput optimization, especially in iSCSI/file based configurations.

Clustering considerations

Protection is necessary for a Tempdb in SQL Server. The Tempdb file is re-created whenever a SQL Server instance is started. In XtremSF, VPLEX, and Cluster Enabler (CE) environments, unshared disk partitions can be used for the Tempdb in a SQL Server 2012 FCI cluster whenever possible to reduce cost and bandwidth.

Considerations for earlier versions

Consider the following when using earlier versions of SQL Server and Windows:

- For SQL Server 2005 and earlier versions, enable the Windows **Instant file initialization** privilege by granting privileges to the SQL Server startup account.
- For versions of Windows earlier than Windows 2008, confirm that the **sector alignment** settings are correct. Windows 2008 and later aligns sectors by default.

General storage considerations

Performance versus capacity considerations

When deploying Microsoft SQL Server, always consider performance, protection, and capacity requirements.

For typical OLTP workloads, throughput measurements in IOPS most likely outweigh the capacity requirement for database files and log files. Tempdb files are typically capacity bound because of the low I/O nature of the workload.

In an OLAP environment, bandwidth measurements in megabytes or gigabytes are more dominant for database files, while Tempdb files are likely to require higher throughput (IOPS).

User database and log files must be well protected to prevent data loss. Because the Tempdb file only contains temporary data and can be re-created when SQL Server starts, the protection for Tempdb is not a high priority. When Tempdb performance is critical (such as in the OLAP environment), it is ideal for Tempdb to use a server flash drive such as XtremSF to minimize storage latency.

When designing storage for different type of workloads, consider the workload type and its typical I/O pattern for database, log, and Tempdb files. Calculate both performance and capacity requirements to ensure that both are satisfied.

Disk type selection

One of the first key decisions you must make when designing SQL Server storage is selecting the type or types of disks that best match your requirements. The types of disks that are appropriate for your SQL Server deployment depend on a number of factors, including your database size and IOPS requirements.

Table 5 shows the disk types that EMC offers with its VNX family of unified storage and Symmetrix VMAX series storage. Flash is also used as XtremSF and XtremSW Cache.

Table 5. Disk types offered by EMC

Disk type	Characteristics	Selection consideration
Fibre Channel (FC)	Reliable disk drives with high read/write speeds.	Ideal for high I/O requirements but might not be appropriate for high capacity requirements
Serial Attached SCSI (SAS)	An improvement over traditional SCSI drives, SAS disks provide high capacity with moderate I/O speed.	Highly suitable for SQL Server environments with high IOPS requirements
SATA	Large capacity disks with less demanding I/O speed.	Appropriate for large databases with low I/O requirements. Most suitable for data warehouse and SharePoint content database
Near-line SAS (NL-SAS)	As with SATA disks, NL-SAS disks are a good fit for less demanding I/O but large capacity requirements.	NL-SAS disks can support large databases at a relatively low cost. NL-SAS disks are typically the best choice for larger databases with low I/O profiles.
Flash	Flash drives have the highest I/O speed with low power consumption.	In general, flash drives can be used as follows: <ul style="list-style-type: none"> • In the storage array as part of the automated storage tiering features, such as EMC FAST VP or FAST Cache to handle any unanticipated I/O spikes • On servers as XtremSF or XtremSW Cache <p>EMC also provides a flash-only array XtremIO™ for the most demanding SQL Server environment.</p>

Follow these general rules when selecting disk types:

- For low IOPS, acceptable disk latency, and high database capacity requirements, use SATA or NL-SAS disks.
- For high IOPS, low disk latency, and high database capacity requirements, use large capacity FC or SAS disks.
- For higher IOPS and very low disk latency requirements but smaller database capacity requirements, use flash drives in storage tiering or FAST Cache.
- For most-demanding IOPS and disk latency requirements but smaller database capacity requirements, use XtremSF, XtremSW Cache.

Different disk types support different IOPS with the same latency requirement. Consider this when calculating disk requirements for your environment. The following table provides random disk IOPS data from the most recent SQL Server validation on EMC VNX and VMAX storage. These results are subject to change based on future testing.

Note: EMC strongly recommends using values from Table 6 when calculating IOPS requirements for SQL Server deployment on VNX and VMAX storage arrays. These numbers will give a baseline of typical acceptable performance as shown in Table 4. For applications that require better performance, add more disks or use array caching such as FAST Cache or server caching such as XtremSW Cache.

Table 6. IOPS for 8 KB random read I/O on various disk types on EMC storage arrays

Disk type	IOPS per disk
15 K rpm SAS	180
10 K rpm SAS	140
7.2 K rpm NL/SAS	70
Solid-state disk (SSD)	3,500

Table 7 lists the IOPS for server flash.

Table 7. SQL Server IOPS for XtremSF models

Random 8K IOPS	XSF550 MLC *	XSF550 MLC *	XSF2200 MLC *	XSF2200 MLC *	XSF320 SLC	XSF700 SLC
Read	131,795	128,207	258,838	256,887	376,072	395,906
Write	23,592	16,235	53,713	35,654	67,635	133,593
R/W (70/30)	56,255	42,471	120,162	93,848	171,666	191,169
* In performance mode. IOPS will be lower when configured in default capacity mode.						

Pools and RAID types

The selection of an appropriate RAID type for your environment is another important decision for a successful implementation of SQL Server. Any RAID type can be used if enough disks are available to handle the I/O and storage capacity requirements. In general, RAID type decisions are based on a set of a given requirements. In order to select an appropriate RAID type for your environment, consider your specific requirements for performance, capacity, and availability.

EMC storage systems support RAID 1/0, RAID 5, and RAID 6 using flash, FC, SAS, NL-SAS, and SATA drives. Each RAID type provides different performance, capacity, and protection levels.

- **RAID 1/0** provides data protection by mirroring data onto another disk. This produces better performance and minimal or no performance impact on in the event of disk failure. In general, RAID 1/0 is the best choice for SQL Server, especially if SATA and NL-SAS drives are used.
- **RAID 5** data is striped across disks in large stripe sizes. The parity information is stored across all disks so that data can be reconstructed. This can protect against a single-disk failure. With its high write penalty, RAID 5 is most appropriate in environments with mostly read I/Os and where large databases are deployed. In the case of SSD flash drives, this performance concern is eliminated and most environments with flash drives can be configured as RAID 5 to support high IO requirements with very low disk latency.
- **RAID 6** data is also striped across disks in large stripe sizes. However, two sets of parity information are stored across all disks so that data can be reconstructed if required. RAID 6 can accommodate the simultaneous failure of two disks without data loss.

Table 8 shows RAID overhead, performance, and storage utilization information for each RAID type.

Note: The RAID overhead value becomes important when performing I/O calculations for the number of disks required. RAID 5 and RAID 6 have an impact on performance when a drive fails and must be rebuilt. In Table 8, the performance is compared using the same number and the same type of disks in the RAID configurations. Storage utilization is compared using the same disk types in the RAID configurations to generate the same IOPs with similar latency.

Table 8. RAID level performance characteristics

RAID level	Random Read	Random Write	Sequential Read	Sequential Write	RAID write overhead value	Storage utilization
RAID 1/0	Excellent	Excellent	Excellent	Excellent	2	Low
RAID 5	Excellent	Moderate	Good	Moderate	4	High
RAID 6	Good	Poor	Good	Moderate	6	Medium

Storage pools are virtual constructs that enable data to move dynamically across different tiers of drives (from high performance to lower cost/high capacity, and vice versa) according to the data's business activity. With VNX and VMAX systems, storage pools are fully automated and self-managing.

The use of storage pools simplifies storage provisioning. Pool-based provisioning provides benefits similar to metaLUNs striping across many drives but, unlike metaLUNs, storage pools require minimal planning and management efforts.

Virtual Provisioning storage considerations

Storage pools support the same RAID protection levels as RAID groups: RAID 5, RAID 6, and RAID 1/0. Multi-tiered pools with different RAID and disk types can be in the same storage pool. Storage pools also provide advanced data services like FAST VP, compression and deduplication, and data protection options such as VNX Snapshots.

Most SQL server database environments can benefit from storage pool based configurations.

EMC VMAX and VNX systems offer virtual provisioning, generally known in the industry as thin provisioning. Thin or virtual provisioning can simplify storage management and reduce storage costs by increasing capacity utilization for many SQL Server use cases.

Virtual provisioning enables SQL Server to acquire more capacity than is physically allocated. The physical storage is allocated to the database “on demand” from a shared pool as it is needed.

Physical storage capacity is fully allocated during LUN creation for thick LUNs. While thin LUNs initially have less physical storage capacity allocated to them, the storage pool provides actual physical storage supporting thin LUN allocations when needed. Physical storage is automatically allocated only when new data blocks are written to the thin LUN.

Both thick and thin LUNs can provide required performance characteristics for any SQL Server workload.

Thin LUN versus thick LUN

Thin devices can be created with inflated capacity, because data devices provide the actual storage space for data written to them. To a host operating system, thin devices have the same configured capacity as standard devices and the host interacts with them in the same way as standard devices.

Thin LUN can be used in most environments with reasonable performance, especially with FAST VP (both VNX and VMAX) and/or with FAST Cache (in VNX).

The main performance considerations related to thin LUNs are:

- Thin LUNs provide significant storage savings and accommodate future growth.
- There is a small performance overhead when a LUN is expanded to accommodate incoming writes.

To summarize, when you determine whether to use storage pools, thin LUNs, thick LUNs, or RAID for LUN configuration, consider the following:

- **Use Storage Pools** to take advantage of data efficiency services (like FAST VP), compression, deduplication, and other pool-based options.
- **Use thin LUNs** with pools for easy setup and management, best space efficiency, energy and capital savings, and databases with flexible space consumption over time.

- **Use thick LUNs** for databases with predictable space requirements.
- **Use RAID groups and traditional LUNs** for databases that do not require changes in size and performance requirements over time, for precisely placing logical data objects on physical drives, and for physically separating data.

Storage sizing best practices

SQL Server performance characteristics can vary substantially from one environment to another, depending on the application. These characteristics fall into two general categories: OLTP generates mostly random read workloads, and data warehouse generates mostly sequential read workloads, as described in [Table 9](#). In an OLTP environment, use read/write IOPS (IO/s) for storage sizing. For a data warehouse environment, use bandwidth (MB/s) for storage sizing.

To get accurate performance estimates, run tests in conditions as close to “real world” as possible. During these tests, use performance monitor logs to capture the characteristics (such as IOPS: reads/second and writes/second, bandwidth: MB/s) of the volumes used to store database files.

Notes:

- Do not use IOPS counters averaged over time as the basis for storage design. EMC recommends identifying the 90th percentile of the IOPS samples and designing for that performance level. This allows the system to respond well to spikes on demand.
 - The IOPS requirements for a RAID group must be calculated independently for reads and writes.
-

Consideration for OLTP database sizing

Sizing for performance

Calculate the number of disks for performance need with the following formula:

$$\text{Number of disks} = \frac{\text{Read IOPS} + (\text{Write IOPS} \times \text{RAID performance overhead})}{\text{IOPS per disk}}$$

Note: You might need to adjust the disk count to conform to the requirements for the selected RAID level. For example, a seven-disk RAID 1/0 set is not possible. In such cases an eight-disk RAID 1/0 set is required.

IOPS per Disk is the number of IOPS of the selected drive type.

[Table 6](#) in [Disk type selection](#) provides the IOPS per disk for different disk types recommended for calculating the disk needs for RAID configuration and tiered storage configuration.

In tiered-storage configuration, you must also calculate the capacity requirements for different tiers. The following tools can help in tiered storage design:

- EMC Storage Configuration Advisor for sizing
You need historical performance data from storage arrays. For more information, refer to the [EMC website](#).
- Workload Performance Assessment Tool
Also known as Mitrend, it shows the FAST VP heat map. For more information, refer to <https://emc.mitrend.com>.
- VSPEX Sizing Tool (VNX)
For more information, refer to <http://www.emc.com/microsites/vspex-ebook/vspex-solutions.htm>.

Sizing for capacity

After the performance sizing is complete, you must consider the storage requirement for capacity. Although typical OLTP database sizing is most likely bound with performance, verify the capacity requirement before the final design is complete.

Calculate the sizing for capacity based on the configuration in Table 9.

Table 9. Sizing for capacity

RAID type	RAID overhead for capacity	RAID overhead for performance (write penalty)	Minimal drives	Minimal disks required for the LUN size
1/0	1/2	2	4	2 × LUN size
5	4/5	4	3	5/4 × LUN size
6	4/6	6	4	6/4 × LUN size

Use the following formula to calculate the number of disks for capacity:

$$\text{Number of disks} = \frac{\text{Total capacity need}}{\text{RAID overhead for capacity}}$$

Note: When calculating the LUN size, consider the future growth of the database size. We recommend that you reserve at least 10 percent of the capacity for database file LUNs.

Final design

Use the larger number as the final sizing to satisfy both performance and capacity requirements.

Best practices for FAST VP sizing

A small percentage of the total utilized capacity is typically responsible for most of the I/O activity in a given database. This is known as workload skew. Analysis of an I/O profile may indicate that 85 percent of I/Os to a volume only involve 15 percent of the capacity. The resulting active capacity is called the **working set**. Software like FAST VP and FAST Cache can keep the working set on the highest-performing tier.

The following are some best practices for sizing FAST VP:

- Ideally, the working set should be in the highest tier (FC/SAS/flash). A typical working set for an OLTP workload is 10 to 30 percent of the database file. Size the top tier in FAST VP to ensure that the working set can be migrated to the highest performance tier.
- FAST VP Performance Tier drives are versatile in handling a wide spectrum of I/O profiles. Therefore, EMC recommends having Performance Tier drives in each pool.
- High capacity drives can help to optimize TCO and often compose 60 to 80 percent of a pool's capacity. Profiles with low IOPS/GB and/or sequential workloads can use a larger number of high capacity drives in a lower tier.
- I/O skew is important to determine prior to sizing FAST VP tiers. SQL skew can vary and depends on the actual SQL Server profile.
- For most VNX sizing tasks, use two tiers only: a performance tier (FC/SAS/flash) and a capacity tier (SATA/NL-SAS). Assume an OLTP workload of 85/15 skew, and adjust according to the actual environment:
 - 85 percent I/O with 15 percent capacity on performance tier (FC/SAS/flash)
 - 15 percent I/O with 85 percent data on capacity tier (SATA/NL-SAS)
- On VMAX, there can be up to three tiers. With an OLTP workload, assume 75/15/10 of hot/warm/cold skew for I/O, then adjust according to the actual environment:
 - 10 percent I/O and 75 percent capacity on SATA
 - 15 percent I/O and 15 percent capacity on FC
 - 75 percent I/O and 10 percent capacity on flash
- EMC Professional Services and qualified partners can assist you in properly sizing tiers and pools to maximize investment. They have the tools and expertise to make specific recommendations for tier composition based on an existing I/O profile.

Sizing for Performance

The IOPS and disk calculation is the same as that given for the RAID group with skew:

Number of Disks for tier =

$$\frac{\text{Read IOPS} + (\text{Write IOPS} \times \text{RAID performance overhead})}{\text{IOPS per disk}} \times \text{I/O skew for tier}$$

Note: Each tier needs to round up to the next logical drive count as the tier with its specific RAID type. For example, for RAID 5 (4+1), a 10-disk set, instead of a seven-disk set, is required.

Capacity-based calculation

Calculate the disk spindles based on the capacity for each tier with the following formula:

$$\text{Number of disks for tier} = \frac{\text{Total capacity} \times \text{capacity skew for tier}}{\text{RAID capacity overhead}}$$

Final design

Use the larger number for each tier as the final sizing to satisfy both performance and capacity requirements.

For details about sizing for FAST VP tiering, refer to the *EMC Virtual Infrastructure for MS Applications enabled by Symmetrix VMAX and MS Hyper-V White Paper*.

Consideration for OLAP database sizing

For predictable bandwidth performance, EMC recommends a building block approach with consideration for the following:

- Targeted bandwidth
- Memory consumption
- CPU number
- Database sizing
- Tempdb sizing

The design principles are as follows:

- **Design for bandwidth, and then for database capacity.**

For OLAP databases, to achieve the goal of completing all the queries in order to generate timely reports, the storage design needs to meet the required bandwidth.

- **Check disk performance.**

For workloads with sequential large read-only I/O (typically 64K or more), calculate the bandwidth for a given I/O size with the IOPS:

$$\text{Number of disks for tier} = \text{Average I/O size} \times \text{read only IOPS}$$

- **Calculate the storage requirement for database files.**

Calculate the number of disks needed for performance as follows:

$$\text{Number of disks for performance} = \frac{\text{Required bandwidth}}{\text{Bandwidth per disk}}$$

Calculate the number of disks needed for capacity as follows:

$$\text{Number of disks for capacity} = \frac{\text{Required capacity}}{\text{RAID overhead for capacity}}$$

Round up to the next logical drive count for its specific RAID type for each of the above calculations.

For tiered storage, calculate tier storage needs for the OLTP workload based on skew, as discussed in [Best practices for FAST VP sizing](#).

- **Use the recommended 1:5 ratio for Tempdb size.**

EMC recommends a 100 GB Tempdb for every 500 GB database file in an OLAP environment. The actual Tempdb size depends on the specific environment.

For details about the OLTP building block design, refer to [Building-block design approach for data warehouse](#). You can also refer to the *SQL SERVER 2012 DATA WAREHOUSE EMC VNX5500 HB, VMware vSphere 5, Windows 2012 White Paper*.

Hypervisor storage considerations

With current server technology rapidly reducing the cost of processing power, most physical server environments become under-utilized, yet some functions must be on separate servers in order for SQL Server applications to work properly. Virtualization can optimize datacenter resources by consolidating server resources down to a few physical servers. This reduces power consumption, saves space, improves return of investment (ROI), increases manageability and flexibility, and introduces new high-availability options.

The Windows Server Virtualization Validation Program (SVVP), which is available on the [Microsoft website](#), provides information about supported EMC storage for [virtualizing SQL Server environments](#).

The virtualization of a SQL Server environment requires some unique storage design best practices.

By virtualizing a SQL Server environment hosted on EMC VNX family or Symmetrix VMAX series storage, customers can use features like VMware vMotion and Microsoft Hyper-V live migration tools. These features enable virtual servers to move between different server hardware platforms without application disruption.

General virtualization guidelines

The following are general guidelines applying to the virtualization of Microsoft SQL Server:

- Follow general SQL Server design principles:
 - Design for performance, reliability, and capacity.
 - Design for user profiles (such as OLTP and OLAP).

- Size virtual machines according to the SQL Server role (part of SAP or SharePoint).
- Physical sizing still applies to calculate the disk number or FAST VP.
- We recommend that you install PowerPath on physical Hyper-V or ESX hosts for load balancing, path management, and I/O path failure detection.
- Virtual machine system resource consideration:
 - Size physical servers to accommodate the number of guests.
 - Keep the number of physical cores and vCPUs in a 1:1 ratio.
 - Ensure that there are no overcommitted CPUs.
 - Do not exceed the size of the NUMA node on the physical server when sizing virtual machines. For details, visit [Using NUMA Systems with ESX/ESXi](#).
 - Fully reserve RAM for SQL Server virtual machines to avoid any memory ballooning.
 - Understand hypervisor limits ([Hyper-V](#) and [VMware](#)):
 - Maximum memory: 1 TB (ESXi 5.1, Windows 2012 Hyper-V)
 - SCSI LUNs per virtual machine: 256 (ESXi 5.1, Windows 2012 Hyper-V)
 - Processor limits
 - VMware vSphere 5.1: 32 vCPUs
 - Windows 2012 Hyper-V: 64 vCPUs
 - [Use large pages in the guest](#) (start SQL Server with a Trace flag—T834) for a dedicated 64-bit SQL Server to improve performance.
 - [Enable Lock Pages](#) in Memory privileges for the SQL Server Service account.
- AlwaysOn Availability Groups considerations:
 - Spread AlwaysOn Availability Groups replicas across multiple physical hosts to minimize potential downtime in the event of physical server issues.
 - If the SQL Server virtual machines in an AlwaysOn Availability Group are part of a hypervisor host-based failover clustering and migration technology, configure the virtual machines to not save/restore their state on the disk when moved or taken offline.
- VMware support for SQL server clusters

There are limitations with both FCI and AAG clusterings for VMware as listed in Table 10.

Table 10. VMware limitations for Microsoft SQL Server clusters

Feature		Shared disk			Non-shared disk	
Microsoft Clustering on VMware		MSCS with shared disk	SQL clustering	SQL AlwaysOn Failover Cluster instance	Network load balance	SQL AlwaysOn Availability Group
vSphere support		Yes	Yes	Yes	Yes	Yes
VMware HA support		Yes ¹	Yes ¹	Yes ¹	Yes ¹	Yes ¹
vMotion DRS support		No	No	No	Yes	Yes
Storage vMotion support		No	No	No	Yes	Yes
MSCS node limits		2 (5.1) 5 (5.5)	2 (5.1) 5 (5.5)	2 ((5.1) 5 (5.5)	Same as OS/app	Same as OS/app
Storage protocols support	FC	Yes	Yes	Yes	Yes	Yes
	In-guest OS iSCSI	Yes	Yes	Yes	Yes	Yes
	Native iSCSI	Yes ²	Yes ²	Yes ²	Yes	Yes
	In-Guest OS SMB	Yes ³	Yes ³	Yes ³	NA	NA
Shared disk	FCoE	Yes ⁴	Yes ⁴	Yes ⁴	Yes	Yes
	RDM	Yes	Yes	Yes	NA	NA
	VMFS	Yes ⁵	Yes ⁵	Yes ⁵	NA	NA

For more details about VMware limitations with Microsoft, refer to [VMware Knowledge Base](#).

- HA/DR consideration: Disable migration technologies that save state and migrate. Always migrate live or completely shut down virtual machines.
- Disable hypervisor-based auto tuning features.

¹ When DRS affinity/anti-affinity rules are used.

² vSphere 5.5 only.

³ Windows Server 2012 Failover Clustering only.

⁴ In vSphere 5.5, native FCoE is supported. In vSphere 5.1 Update 1 and 5.0 Update 3, two node cluster configuration with Cisco CNA cards (VIC 1240/1280) and driver version 1.5.0.8 is supported on Windows 2008 R2 SP1 64-bit Guest OS. For more information, see the VMware Hardware Compatibility guide: [Cisco UCS VIC1240](#), [Cisco UCS VIC1280](#)

⁵ Supported in Cluster in a Box (CIB) configurations only. For more information, see the [Considerations for Shared Storage Clustering](#) section in this article.

For optimal performance with Tier 1 mission critical SQL server instances:

- Follow the same guidelines as for a physical environment; separate LUNs for data and logs.
- Ensure that each ESXi or Hyper-V server has at least four paths (two HBAs) to the storage, with a total of four ports. Ensure that connectivity to the storage array is provided to both SPs for VNX, or across multiple front-end directors on VMAX.
- Place SQL Server storage on separate disks from guest OS (VHD/VHDX or VMDK) physical storage.

For acceptable performance with Tier 2 and lower service level instances:

- Virtual Hard Drives for VM OS LUNs can share a LUN at the hypervisor level (datastore for vSphere or the host level LUNs for VHD/ VHDX in Hyper-V) with the SQL Server database and log LUNs if ease of deployment is the main concern and performance is secondary.
- Multiple virtual machines or databases can share LUNs at the hypervisor level (datastore for vSphere or the host level LUNs for VHD/ VHDX in Hyper-V) if the HA/DR level is acceptable for the specific environment.

Microsoft provides additional information and recommendations for virtualizing SQL Server.

Best practices for the VMware vSphere environment

The following are some general best practices for deploying SQL Server in a VMware vSphere virtual environment:

- Deploy application virtual machines on shared storage. This allows the use of vSphere features like vMotion, HA, and DRS.
- Create VMFS file systems from vCenter to ensure partition alignment.
- Add about five percent CPU requirements for hypervisor overhead.
- When using VMFS for data storage, format VMDK files as **eagerzeroedthick** (specifically for database and log files.)
- Use multiple PVSCSI adapters and evenly distribute target devices.

VMware Paravirtual SCSI (PVSCSI) adapters are high-performance storage drivers that can improve throughput and reduce CPU use. PVSCSI adapters are best suited for SAN environments, where hardware or applications drive high I/O throughput.

Figure 5 shows that the SCSI controller type was changed to Paravirtual to improve driver efficiency. The default SCSI controller driver is LSI Logic SAS. LUNs are spread across all available SCSI drivers.

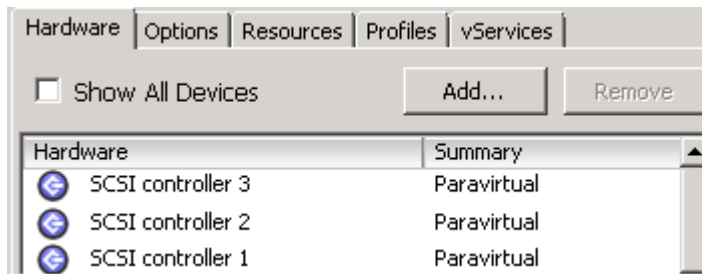


Figure 5. SCSI controllers

- Increase the queue depth of HBA in the ESXi server to improve the throughput of HBA cards on ESXi servers. We observed a benefit in setting the queue depth of the HBA to 64 (the default is 32 in ESXi 5.0, and 64 in ESXi 5.1). Results may vary in specific environments.

Notes:

- The highest queue depth configurable for the HBA ports in ESXi 5.0 is 128 and ESXi 5.1 is 256, if associated LUNs have dedicated ESXi server ports.
- For ESXi 5.0 servers with multiple virtual machines, the value of Disk.SchedNumReqOutstanding in VMware’s advanced options needs to match the queue depth.

For more information about ESXi server configuration, refer to the [VMware knowledge base website](#) and [VMware ESX scalable storage performance manual](#).

VMFS versus RDM

VMFS and RDM generally have similar performance, with RDM providing slightly better performance. VMFS can be used in most environments unless there is a specific requirement for clustering and snapshot replication. Refer to the [VMware website](#) for details.

Table 11 compares VMFS and RDM in the VMware environment for SQL Server.

Table 11. VMFS versus RDM

VMFS	RDM
Better storage consolidation—multiple virtual disks and virtual machines per VMFS LUN; but still supports one virtual machine per LUN	Enforces 1:1 mapping between virtual machine and LUN
Consolidating virtual machines in LUNs—less likely to reach the ESX LUN limit of 255	More likely to hit the ESX LUN limit of 255
Managing performance: combined IOPS of all virtual machines in the LUN is less than the IOPS rating of the LUN	Not impacted by IOPS of other virtual machines
Limited LUN size (2 TB for vSphere 5.1 and 62 TB for vSphere 5.5)	Unlimited LUN size

VMFS	RDM
Works well for most environments other than those requiring raw device mapping (RDM)	Required for clustering (Quorum disk); for example, SQL Failover Clustering Required for SAN management tasks such as backup and snapshots

Microsoft Hyper-V

Storage options for SQL Server

VHDX is the new virtual hard disk format introduced in Windows Server 2012 Hyper-V; it provides increased storage capacity and data protection.

VHD, the Windows Server 2008 Hyper-V virtual hard disk, can be converted to VHDX in Windows Server 2012 Hyper-V.

SQL Server database files and log files can reside on VHD or VHDX. VHDX can be created on a Cluster Shared Volume (CSV) to take advantage of the HA/DR features that Hyper-V provides.

CSV is a feature of failover clustering, which was first introduced in Windows Server 2008 R2 for use with Hyper-V. CSV is a shared disk containing an NT file system (NTFS) volume that is made accessible for read and write operations by all nodes within a Windows Server Failover Cluster.

PassThrough disk can bypass the hypervisor layer and maximize the performance of the underlying disks.

NPIV (N_Port ID Virtualization) is a Hyper-V virtual Fibre Channel feature introduced in Windows 2012. It allows connection to Fibre Channel storage from within a virtual machine. With NPIV, virtualized workloads can use existing Fiber Channel investments. It also supports many related features, such as virtual SANs, live migration, and Multipath I/O (MPIO).

This feature provides close to pass-through disk performance with Hyper-V level protection and easy migration.

Virtual Fibre Channel for Hyper-V provides the guest operating system with unmediated access to a SAN by using a standard World Wide Name (WWN) associated with a virtual machine. Hyper-V users can now use Fibre Channel SANs to virtualize workloads that require direct access to SAN LUNs. Fibre Channel SANs also allow you to operate in new scenarios, such as running the Failover Clustering feature inside the guest operating system of a virtual machine connected to shared Fibre Channel storage.

EMC storage arrays are capable of advanced storage functionality that helps offload certain management tasks from the hosts to the SANs. Virtual Fibre Channel presents an alternate hardware-based I/O path to the Windows software virtual hard disk stack. This allows you to use advanced functionality like hardware snapshots directly from Hyper-V virtual machines. For details about Hyper-V virtual Fibre Channel, refer to [Microsoft TechNet](#).

SQL Server clustering storage considerations

AlwaysOn Availability Groups (AAG), a high availability and disaster recovery feature introduced in SQL Server 2012, requires Windows Server Failover Clustering (WSFC). AlwaysOn Availability Groups are not dependent on a SQL Server Failover Clustering Instance (FCI).

In a SQL Server Failover Clustering Instance (FCI), the SQL Server database and log files are shared among all the nodes in the cluster, so the storage LUNs hosting these files need to be accessible from all nodes. This means that all the LUNs need to be configured and zoned to the nodes in the cluster simultaneously. The specific database/log LUNs can only be accessed from the node actively running the SQL Server instance.

While primary and secondary copies of databases are not sharing storage in AlwaysOn Availability Groups (AAGs), each node in the cluster needs to have its own storage configured and zoned to it. The database copy on the secondary node of an AlwaysOn Availability Group can be accessed if it is configured as a “readable copy,” which is independent of the primary copy. This can be used to enable reporting capability on the secondary copy to offload from the primary copy.

Figure 6 shows the difference between FCI and AAG.

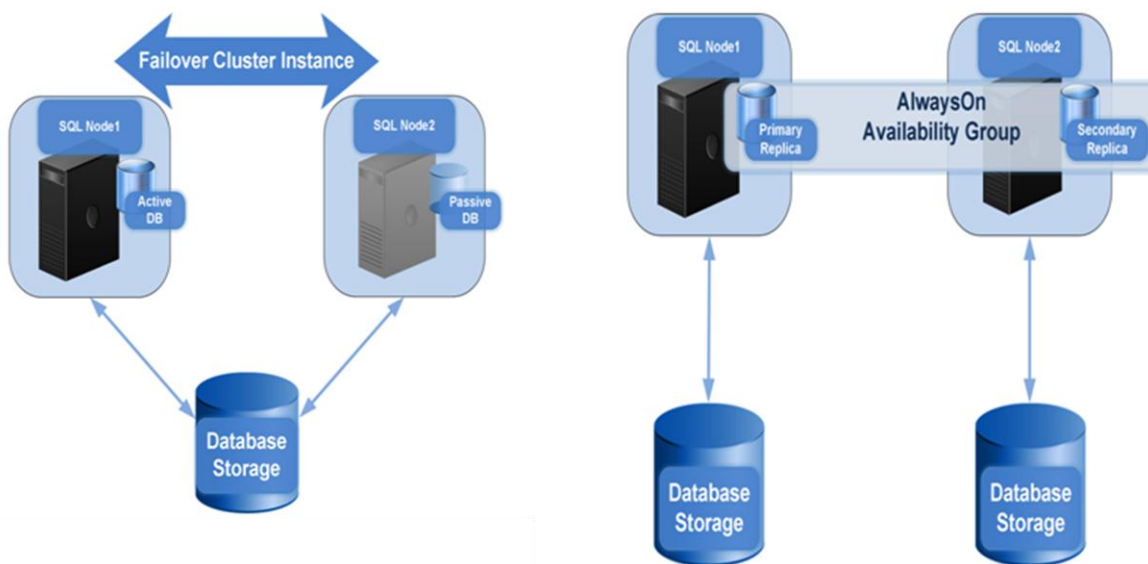


Figure 6. FCI versus AAG

The limitations for SQL Server clustering support in a VMware environment are detailed in [Table 10](#).

Symmetrix VMAX storage design guidelines

The EMC Symmetrix VMAX series is high-end storage for the data center. The system scales to a massive 2 PB and consolidates more workloads with a much smaller footprint than other arrays. EMC Symmetrix Virtual Matrix Architecture seamlessly scales performance, capacity, and connectivity on demand to meet all application requirements. The system supports flash drives, Fibre Channel, and SATA drives, as well as optimized automatic tiering with FAST VP. The system also supports virtualized and physical servers, including open systems, mainframe, and IBM i servers.

VMAX series hardware design considerations

Some of the most important design considerations for SQL Server on VMAX are listed below:

- When creating LUNs, use fewer but larger hypervolumes to improve performance.
- Use a minimum of two HBAs per server, with each HBA connected to at least two director ports (across multiple VMAX engines, if possible).
- For thick and thin LUNs, use striped metavolumes.

Virtual Provisioning considerations and best practices

With Microsoft Windows “thin-friendly” NTFS formatting capabilities and Microsoft SQL Server’s Instant File Initialization mechanisms, SQL Server can derive the maximum benefits of Virtual Provisioning with EMC Symmetrix. Thus, thin LUN pools are recommended for SQL Server on Symmetrix VMAX. Thin device performance is equivalent to regular (thick) device performance on VMAX, and in most cases, the use of thin pools can reduce the initial storage requirement.

The following summarizes best practices when configuring Microsoft SQL Server with Virtual Provisioning on EMC Symmetrix VMAX:

- Use Virtual Provisioning when system over-allocation of storage is typical.
- Use Virtual Provisioning when expecting rapid growth over time but downtime is limited.
- Configure SQL Server databases with Instant File Initialization (this is the default for SQL Server 2012).
- Consider using the thin pool utilization threshold tool to monitor the pools and avoid thin pools running out of space.

Avoid Virtual Provisioning for the following environments:

- Systems where shared allocations from a common pool of thin devices do not meet the customer requirements.
- Systems where large amount of deleted space cannot be reclaimed.
- Systems that cannot tolerate an occasional response time increase of approximately 1 millisecond because of writes to uninitialized blocks.

FAST VP considerations and best practices for a VMAX storage system

FAST VP provides SQL Server with reduced administration and faster space and I/O issue resolution. Especially with OLTP workload, FAST VP provides efficient storage usage and moves the most frequently accessed data to the highest performing tier.

The following are some considerations and best practices when using FAST VP:

- Avoid putting log LUNs in a flash tier because logs generate mostly sequential writes and will not benefit from a flash device.
- Bind database LUNs to the FC tier to start with FAST VP and provide enough storage to accommodate the capacity if the hot data needs to move to the upper tier.

- When using FAST VP with AlwaysOn Availability Groups, place availability group copies of the same database in different pools to achieve better availability.
- Logs can be pinned in a specific tier when placed in the same storage pool with data file LUNs.
- Perform sizing according to [Best practices for FAST VP sizing](#).

Note: The skew ratios can vary and depend on the specific SQL Server profile.

VNX storage design guidelines

The EMC VNX family delivers industry-leading innovation and enterprise capabilities for file and block storage in a scalable, easy-to-use solution. This next-generation storage platform combines powerful and flexible hardware with advanced efficiency, management, and protection software to meet the demanding needs of today's enterprises.

The VNX family includes the VNXe series, purpose-built for the IT manager in entry-level environments, and the VNX series, designed to meet the high-performance, high-scalability requirements of midsize and large enterprises.

EMC FAST Suite is an advanced software feature providing greater flexibility to manage the increased performance and capacity requirements of the SQL Server environment. EMC FAST suite makes use of SSD drives, SAS, and NL-SAS storage configuration to balance performance and storage needs. FAST suite includes FAST Cache and FAST VP.

Application Protection Suite automates application-consistent copies and enables you to recover to defined service levels. User roles enable self-service copy management, while improving visibility for all application recovery points. Alerts are generated automatically, providing fast resolution to recovery gaps. Integrated reporting can prove compliance with protection policies. Applications supported include Oracle; Microsoft SQL Server, SQL Server, and SharePoint; VMware; and Hyper-V. Application Protection Suite includes:

- For VNX series: Replication Manager, AppSync, and Data Protection Advisor for Replication Analysis
- For VNXe series: Replication Manager

FAST Cache considerations and best practices

EMC FAST Cache increases the storage system cache by extending the functionality of DRAM cache, mapping frequently accessed data to SSD. FAST Cache capacities range from 73 GB to 2 TB, which is considerably larger than the available DRAM cache of the existing storage systems. If a particular chunk of data is accessed frequently by the user application, that chunk will be automatically promoted into the FAST Cache by copying it from hard disk drives to the flash drives. Subsequent access to the same chunk is serviced at flash drive response times, thus boosting the performance of the storage system.

FAST Cache is most suitable for I/O-intensive random workloads with small working sets. A typical OLTP database with this kind of profile can greatly benefit from FAST

Cache to improve performance and response time. Monitor SQL Server database filegroups and enable FAST Cache on the highly active storage pools where such data is located.

The *EMC FAST Cache: a Detailed Review*, available on the EMC Online Support website, provides more details on FAST Cache design concepts, planning, and usage guidelines.

Testing suggests that the inclusion of FAST Cache results in a 300 percent increase in transactions per second (TPS) for a SQL OLTP workload using the same number of hard disk drives in the back end.

The *EMC Unified Storage for Microsoft SQL Server 2008: Enabled by EMC CLARiiON and EMC FAST Cache Reference Architecture*, available on EMC Online Support, provides more details on how to build a solution with EMC FAST Cache.

When using FAST Cache, allow sufficient time for the cache to warm-up to fully utilize the cache. In our tests with OLTP workloads, FAST Cache took about 1 to 2 hours to warm-up.

The warm-up time depends on the type and number of back-end hard disk drives, the size of FAST Cache, and the size of the working set. EMC Unisphere has several FAST Cache counters that you can monitor to obtain optimum utilization and warm-up time. The *EMC CLARiiON, Celerra Unified, and VNX FAST Cache White Paper*, available on EMC.com, provides more information.

FAST VP considerations and best practices

According to industry analysts, 60 to 80 percent of operational database data is inactive and increases as the size of the database grows. Low-cost, high-capacity spinning drives are an ideal choice for inactive data, while high performance drives are suitable for frequently accessed data. Manually classifying and storing data in the right tier is a complex and challenging task, and usually requires down time to move data.

EMC FAST VP moves frequently accessed data to a faster physical storage within a pool and moves infrequently accessed data to a less-expensive physical storage within the pool automatically.

Using FAST VP you can:

- Control when FAST VP can move data to prevent any impact on host I/O requests during known periods of high system usage.
- Improve overall system performance without investing in additional high-performance physical drives.
- Apply FAST VP to any or all pool-based database LUNs on a storage system.
- When log LUNs need to be in the same pool of database LUNs, pin them to the SAS tier and disallow data relocation for these LUNs.
- In an OLTP workload, pin the Tempdb in the SAS tier and disallow data relocation for these LUNs.
- Set the FAST policy to Start High then Auto Tier (default).

Understanding Microsoft SQL Server OLTP Performance Characterization for EMC VNX Series, available on the EMC Online Support website, provides more details about using FAST VP with SQL Server.

FAST VP operates in the background, while the LUNs are online and available for host access. The data movement speed can be controlled to minimize the impact on overall system performance (you can configure the relocation rate to high, medium, or low).

FAST Cache versus FAST VP

FAST Cache boosts the performance immediately for random-access patterns and burst prone data, while FAST VP allows the storage system to move inactive data to a lower-cost storage tier. FAST Cache operates in 64 KB units while FAST VP operates in 1 GB chunks. FAST VP can promote both thin metadata and the most-accessed user data to the higher tier, while FAST Cache promotes the metadata to improve thin LUN performance.

FAST Cache and FAST VP can work together in a SQL Server database to improve performance, achieve higher capacity utilization, and lower the power and cooling requirements.

The following are best practices when using FAST Cache and FAST VP:

- When a limited number of Flash drives is available, use flash drives to create FAST Cache first.
 - FAST Cache is global and can benefit multiple pools in the storage system.
 - FAST Cache uses 64 KB chunks while FAST VP uses 1 GB chunks, which results in higher performance benefits and faster reaction time for changing usage patterns.
- Use flash drives to create a FAST VP performance tier for a specific pool to ensure the performance of mission-critical data. FAST VP tier is dedicated to a storage pool and is not shared with other storage pools in the same storage array.

Server flash considerations

XtremSF overview

XtremSF is a single, low-profile server flash hardware card that fits in any rack-mounted server within the power envelope of a single PCIe slot, available with a broad set of eMLC and SLC capacities. It can be deployed:

- As local storage that sits within the server to deliver high performance
- In combination with XtremSW Cache server-caching software to improve network storage array performance, while maintaining the level of protection required by critical application environments

Design best practices for XtremSF

SLC and MLC NAND XtremSF offer capabilities that serve two different types of applications—those requiring high performance at an attractive cost per bit (MLC) and those that are less cost sensitive and seeking higher performance over time (SLC).

In a high-demand data warehouse or OLAP environment, and sometimes in an OLTP environment where Tempdb is heavily utilized, XtremSF can be used as Tempdb storage to reduce Tempdb contention and thus improve the bandwidth.

XtremSF is best for SQL Server databases with a 70 to 90 percent read/write ratio and with SQL Server-level data protection. With larger XtremSF cards, it is possible to fit an entire database into a single XtremSF card.

Use cases that can benefit from local XtremSF are detailed in the *EMC XtremSF Performance Acceleration for Microsoft SQL Server 2012 White Paper*.

XtremSW Cache overview

XtremSW Cache is EMC server-caching software for flash PCIe cards that populates XtremSF with data so it can be used as cache.

XtremSW Cache is designed to follow these basic principles:

- **Performance:** Reduce latency and increase throughput to dramatically improve application performance.
- **Intelligence:** Add another tier of intelligence by extending FAST array-based technology into the server.
- **Protection:** Deliver performance with protection by using the high availability and disaster recovery features of EMC networked storage.

SQL Server workloads that can benefit most from XtremSW Cache are:

- Applications that have high read-to-write workload ratios. Maximum effectiveness is gained where the same chunks of data are read many times and seldom written.
- Applications with a small working set, which receive the maximum possible boost.
- Applications with predominantly random read workloads. Sequential workloads that tend to have a significantly larger active dataset (like data warehousing) in proportion to the available XtremSW Cache size benefit little from XtremSW Cache.
- Applications with a high degree of I/O concurrency (that is, multiple I/O threads).
- Applications with smaller I/O sizes (8 KB or lower) but generate large I/O sizes. The XtremSW Cache software enables you to tune features such as page size and maximum I/O sizes, which helps in these environments to accelerate particular I/Os and ignore others (like backup read I/Os).

XtremSW Cache can accelerate read operations, while all write operations are written to the storage array and are not affected by XtremSW Cache. In many cases, improvement in write-throughput performance can be observed as XtremSW Cache offloads the read operations, enabling the array to handle more write operations as a side benefit. XtremSW Cache may not be suitable for write-intensive or sequential applications such as data warehousing, streaming, media, or big data applications.

To summarize, you can use XtremSF as local storage for read and write acceleration, temporary data, and large working sets, while XtremSF with XtremSW Cache can be used for read acceleration of mission-critical data with small working sets that require data protection.

Design best practices for XtremSW Cache

Working from the base storage configuration, determine which SQL Server needs XtremSW Cache acceleration.

In a typical SQL Server OLTP environment:

- Use XtremSF with XtremSW Cache for read acceleration of mission-critical data with working sets small enough to fit in the cache.
- Use XtremSW Cache Performance Predictor for initial benefit analysis of the SQL Server with XtremSW Cache.
- The read-intensive database data file LUNs generally have heavy workloads subjected to a high-read skew and are good candidates for XtremSW Cache.
- SQL Server OLTP data files experience constant random reads and contribute to the overall duration of transaction times. Data files also experience regular bursts of write activity during a checkpoint operation. Using XtremSW Cache to cache reads and avoid an I/O workload on the EMC array enables the array to consume burst writes faster and avoid any read delays for transactions.
- Log LUNs and Tempdb LUNs in OLTP databases are write-intensive and typically do not benefit from XtremSW Cache.
- In SQL Server AlwaysOn environments, the secondary databases do not need to be accelerated unless a specific performance requirement justifies the use of XtremSW Cache.
- Set the page size to 64 KB in the XtremSW Cache to accommodate large I/O for the SQL Server database.
- If the workload is not expected to increase after deploying XtremSW Cache, there is no need for additional system resources such as memory or CPU.
- Have at least two XtremSF cards within the server infrastructure when redundancy is required.
- Deduplication is generally not beneficial to SQL Server's I/O pattern.

Design best practices for XtremSW Cache in a virtualized environment

The following are some design best practices for XtremSW Cache in a virtualized environment:

- Have at least two XtremSF cards within your hypervisor server infrastructure when redundancy is required.
- Where vMotion is required, calculate the XtremSF capacity and placement so that the remaining server and XtremSF capacity still have the ability to serve the configured XtremSW Cache settings of all virtual machines during a vMotion operation.

For example, if 10 virtual machines are configured to use 100 GB XtremSW Cache, which requires a total of 1 TB XtremSW Cache capacity, during a vMotion operation, the remaining servers in the virtualized cluster with XtremSW Cache must facilitate at least 1 TB of cache space.

- If applications only need a small part of the XtremSF card capacity for each virtual machine, the virtual machines with these applications can share the same physical card and are best placed on the same ESXi or Hyper-V host.
- In cases where, a certain application demands all the available capacity of the XtremSF card, the host should dedicate that specific card to the virtual machine.
- Multiple XtremSF cards can be installed on the same server, if required.
- Multiple XtremSF cards can be configured to the same hypervisor to create multiple cache devices for the virtual machine.
- For specific application workloads that have been selected to use the split-card feature, part of the card can be configured to serve the caching needs of the virtual machine; the other part can be configured as XtremSF storage to serve the need for a temporary datastore, such as a Tempdb storage space.
- Additional considerations for the portability of the virtual machine in the virtualized environment are necessary for this configuration because the virtual machine now depends on the storage that is local to that server.
- The minimum size for the XtremSW Cache vDisk is 20 GB for any virtual machine that needs flash cache acceleration.
- Placing only the VHDXs that require XtremSW Cache acceleration on the LUNs configured with XtremSW Cache. All VHDXs on the LUN configured with XtremSW Cache will be accelerated.

Sizing consideration for XtremSF and XtremSW Cache

Sizing recommendations are available for each application type. Environments are different so the implementations are different. The following are the minimum configurations recommended for each application, based on our testing with a typical database workload and the application workload. In most cases, adding more XtremSW Cache gives better performance until the size of the cache is equal to or greater than the working set.

To determine the sizing that best fits a specific application and environment, consider both the level of performance you need and the cost.

Table 12 provides XtremSW Cache recommendations for each application. The cache-to-storage ratio⁶ largely depends on the active working set of the database and will change based on actual usage. These recommendations are based on our testing in a controlled environment. Your environment might be different, so use the numbers provided as a guideline.

⁶ XtremSW Cache-to-storage ratio is the cache and database storage size ratio. If the ratio is 1:10, then for each 10 GB of data, provide at least 1 GB of XtremSW Cache.

Table 12. Recommended cache for each application

Application	Database type	Read-to-write ratio	Recommended XtremSW Cache-to-storage ratio
SQL Server	OLTP	90:10	1:10
SQL Server	OLTP	70:30	1:5
SharePoint	Content/crawl	100% read	1:5
SQL Server	OLAP	100% read	1:5 (Tempdb on XtremSF: database size)

For SQL Server OLAP applications, such as a data warehouse environment, eMLC XtremSF (alone or in split-card mode) can be used as the Tempdb to improve query performance. Consider at least 200 GB Tempdb space for every 1 TB of database space.

For details of all design and configuration best practices regarding XtremSF and XtremSW Cache, refer to the *EMC VSPEX with EMC XtremSF and EMC XtremSW Cache Design Guide*. This document is for VSPEX, but the design principles and best practices can apply to most environments.

Automation with ESI

EMC Storage Integrator (ESI) for Windows Suite is a set of tools for Microsoft Windows and Microsoft applications administrators. The suite includes ESI for Windows, ESI PowerShell Toolkit, ESI Service, ESI Management Packs for System Center Operations Manager (SCOM), and ESI Service PowerShell Toolkit.

- **ESI for Windows**

Provides the ability to view, provision, monitor, and manage block and file storage for Microsoft Windows. ESI supports the EMC Symmetrix VMAX series and EMC VNX series. ESI also supports storage provisioning and discovery for Windows virtual machines running on Microsoft Hyper-V, Citrix XenServer, and VMware vSphere.

As shown in Figure 7, ESI simplifies the storage management for Windows and automatically applies some best practices with storage configuration thus simplifying storage deployment for Windows.

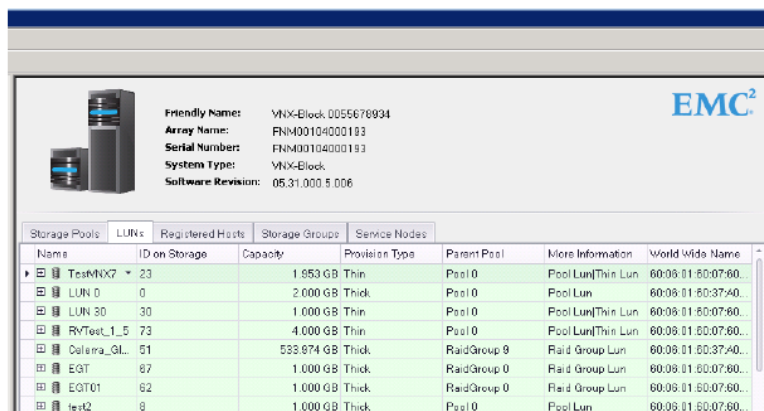


Figure 7. ESI for Windows storage management

- **ESI Management Packs for Systems Center Operations Manager 2012**

Enable storage infrastructure management in a single user interface. With the SCOM 2012 integration, Windows administrators can discover storage assets, map physical objects and logical objects, manage alerts, and roll-up health states, with configurable parameter thresholds from within System Center 2012.

- **EMC PowerShell Toolkit (PSToolkit)**

Is a powerful utility designed for administrators and users of Windows to assist in storage system management. PSToolkit cmdlets enable storage system administrators to get storage system information; create and delete storage pools, storage groups, and storage volumes; and map and mask these to available host servers. This toolkit enables administrators to efficiently create automated scripts for dynamic creation and deletion of virtual machines as needed by users.

SQL Server protection

Overview

Microsoft enhanced its native SQL Server high availability and data protection capabilities at the database level for SQL Server 2012 by introducing the AlwaysOn Availability Groups feature. EMC has a number of data protection products and options that complement AG and can further protect your SQL Server environment from the loss of a database, server, or an entire site. Various SQL Server 2012 high availability and disaster recovery options are described in this section.

EMC storage systems offer a wide range of features for SQL Server database protection and high availability. EMC replication technologies, such as TimeFinder®, SRDF and VNX snaps/clones, provide best data protection in the industry. RecoverPoint with continuous protection and Replication Management tools, such as AppSync and Replication Manager, provide protection for SQL Server in the application level.

You can leverage EMC technology in the SQL Server backup processes to:

- Reduce production system impact during the backup processing.
- Create consistent backup images.
- Integrate SQL Server backup and recovery processes.

AlwaysOn Availability Groups

In the event of a hardware or software failure, multiple database copies in an AlwaysOn Availability Group enable high availability with fast failover and no data loss. This eliminates end-user downtime, which is a significant cost to recover a past point-in-time backup from a disk or tape, as shown in Figure 8. AlwaysOn Availability Groups can be extended to multiple sites and can provide resilience against data center failures. It provides database level replication with automated failover.

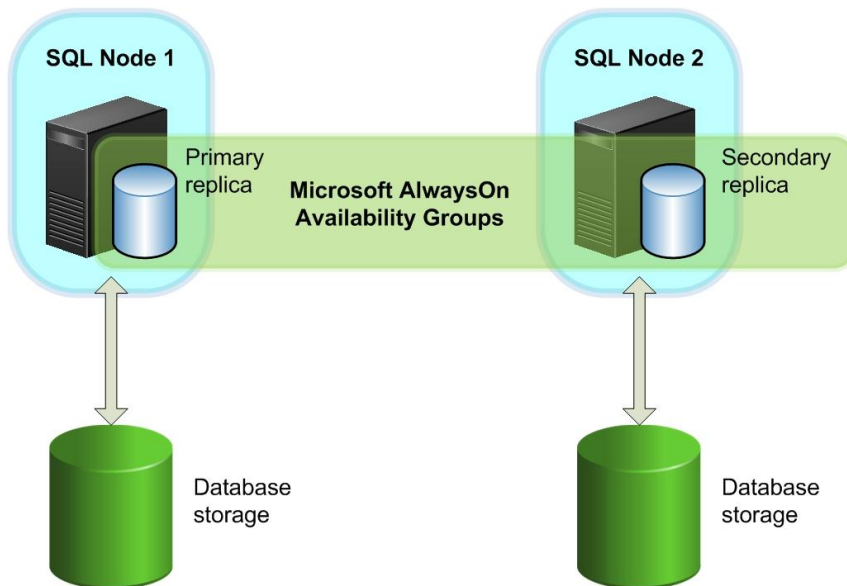


Figure 8. SQL Server AlwaysOn Availability Groups

SQL Server native data protection

If a past point-in-time copy of a database is required, you can use SQL Server to create a lagged copy in an AG environment. This can be useful if a logical corruption replicates across the databases in the AlwaysOn Availability Group, which results in the return to a previous point in time. This is also useful if an administrator accidentally deletes the user data. EMC has the ability to provide the same or better protection levels but use far less storage with the use of snapshots.

Recoverable versus restartable copies

EMC replication technologies can create two different types of database copies, recoverable or restartable. You can use either or both technologies to meet the needs of backup RPO and other requirements.

Recoverable database copy

A recoverable database copy is a backup in which logs are applied to the database and rolls forward to any point in time after the database copy was created. In the event of a production database failure, the database can be recovered not only to the point in time of the last backup, but can also roll forward the subsequent transactions up to the point of the failure. This is a very important feature for the SQL Server database as well as for many other business requirements. Table 13 lists the three ways of creating a recoverable copy of a SQL Server database.

Table 13. SQL Server methodology to create a recoverable database copy

Backup methodology	Description	Supported
Stream backup	T-SQL statement or native SQL Server backup	Native SQL backup, Networker®
VDI	Virtual Device Interface for third party software	VMAX/VNX/VNXe Networker, Replication Manager, AppSync, RecoverPoint
VSS	Volume Shadow Copy Service for third-party software	VMAX/VNX/VNXe Networker, Replication Manager, AppSync, RecoverPoint

All three types of backup are integrated with SQL Server and are considered to be hot backups. SQL Server records the backup when it takes place.

Restartable database copy

When a database copy is created in a storage level without any database-level integration, SQL Server can use the database copy to perform a crash recovery and bring the database to the point in time that copy was taken. This is considered a restartable database copy.

In this case, all transactions recorded as committed and written to the transaction log, with its corresponding data pages written to the data files, are rolled forward (redo). SQL Server will then undo or roll back changes recorded but never committed (such as dirty pages flushed by a lazy write).

This results in a database state with a transactionally consistent point in time. Additional transaction log backup cannot be applied to a database in this state, thus making the database recovered only to the point in time of the backup.

A restartable database copy is considered as a cold copy of the database. There is no record of backup in SQL Server.

EMC Consistency technology can be used to create restartable copies of a source SQL Server database. This type of technology is noninvasive to production database operations and occurs entirely at the storage array level. The restartable images represent dependent-write consistent images of all related objects that have been defined within the consistency group.

VDI and VSS frameworks for backup replication

EMC VMAX, VNX, and VNXe also implement consistency technology integrated with VDI snapshot technology and VSS framework to create a recoverable database copy.

EMC RecoverPoint, Replication Manager, and AppSync are built on top of these technologies to provide data protection to suit different environments.

EMC high availability and data protection offerings for SQL Server

While SQL Server native data protection is sufficient for some customers, most still require full backup and restore capabilities for SQL Server databases. EMC offers a wide range of options to provide high availability and data protection with SQL Server. [SQL Server protection solution](#) in Appendix D provides solution details including some of the listed protection offerings for SQL Server. Table 14 lists the SQL Server database options.

Table 14. EMC high availability and data protection options

Category	Tool/System	Features	Description
Continuous availability	RecoverPoint	CDP	<ul style="list-style-type: none"> Synchronous Local recovery protection
		CRR	<ul style="list-style-type: none"> Asynchronous Continuous remote replication
		CLR	<ul style="list-style-type: none"> Concurrent local and remote data Combines CDP and CRR
	VMAX/VNX with a built-in RecoverPoint splitter	CDP/CRR/CLR	Both VMAX and VNX arrays have options with a built-in RecoverPoint splitter that functions as native continuous availability
	VMAX	SRDF	Continuous replication
Point-in-time rapid replication recovery	AppSync	Snapshot only replication on VNX	<ul style="list-style-type: none"> A simple, service-level agreement (SLA)-driven, self-service data protection, storage management for SQL Server Also works with RecoverPoint on VNX No agent necessary

Category	Tool/System	Features	Description
	Replication Manager	Snapshot/Clone, SAN copy for VMAX and VNX	<ul style="list-style-type: none"> • A comprehensive data protection software • Must install agent on SQL Server
	VMAX TimeFinder	Mirror	General monitor and control operations for business continuance volumes (BCVs)
		CG	Consistency groups
		Clone	Clone sessions generally consume the same size of production LUNs, but have no impact once created
		Snap	Snapshots consume less space than clones, but have more impact on production LUNs if the data changes frequently on the LUNs
	VNX	Clone	Clone sessions generally consume the same size of production LUNs, but have no impact once created
		Snap	Snapshots consume less space than clones, but have more impact on production LUNs if the data changes frequently on the LUNs
Point in time efficient backup and restore	EMC Avamar®	Complete software and hardware solution	Variable-length deduplication significantly reduces the backup time by only storing unique daily changes while maintaining full daily backups for immediate, single-step restore
	EMC Networker	Traditional backup and restore software solution	Centralizes, automates, and accelerates data backup and recovery with a wide range of data protection options

Each product has its own benefits and considerations. The decision depends on the service-level requirements of each use case.

EMC hardware-based snap and clone products have been integrated with Microsoft VDI and VSS technology for many years. Symmetrix TimeFinder and VNX SnapView (or advanced snap in the later version) enable local point-in-time snapshots and data clones for backup and recovery operations. These products enable simple, non-destructive backup operations with space-saving snapshots or full block-for-block clone copies of your databases and logs. With these products, backups and restores can occur in seconds.

EMC Replication Manager enables the management of EMC point-in-time replication technologies for SQL Server through a centralized management console. Replication Manager coordinates the entire data replication process—from discovery and configuration to the management of multiple, application-consistent, disk-based replicas. The databases can be automatically discovered with streamlined management for replication scheduling, recording, cataloging, and auto-expiration.

EMC strongly recommends a robust method that enables rapid SQL Server database backup and restore. EMC Replication Manager, EMC Avamar, and EMC Networker offer features for log truncation and the mounting of databases to alternative hosts.

Even if the native Microsoft SQL Server 2012 AAG is used, EMC strongly recommends an alternative, solid, point-in-time SQL Server data protection strategy to guard against logical corruption events.

Replication technologies

SQL Server requires dedicated system resources. When a protection mechanism is implemented in the SQL environment, consider the performance impact on SQL Server that comes with it.

EMC RecoverPoint

RecoverPoint replicates data to protect the SQL Server environment from disaster. It provides three options:

- **Local recovery protection (CDP)** provides synchronous protection by capturing every transaction in a database and simultaneously writing it to a secondary storage location.
- **Continuous remote replication (CRR)** is an asynchronous protection that can replicate data across any distance.
- **Concurrent local and remote data protection (CLR)** that combines the CDP and CRR replication methods to provide local and remote protection for a SQL Server environment.

RecoverPoint scales well, and can be implemented on very large SQL Server environments. It can provide continuous restartable copy of user database and log files to almost any point in time.

Newer VNX and VMAX arrays have RecoverPoint splitter build in. *EMC RecoverPoint Replicating Microsoft SQL Server Technical Notes*, available from the EMC Online Support website, provides more information.

EMC TimeFinder

EMC TimeFinder is business continuity features that allow control operations on device pairs within a local replication environment with the following functionality:

- TimeFinder/Mirror—General monitor and control operations for business continuance volumes (BCV)
- TimeFinder/CG—Consistency groups
- TimeFinder/Clone—Clone copy sessions
- TimeFinder/Snap—Snap copy sessions

These features can be used by storage manager or coupled with other data replication software to provide SQL Server integrated VDI or VSS copies of snap and clones.

VNX clone/snap

Local Protection Suite combines snapshots and clones with point-in-time recovery with DVR-like rollback capabilities for business continuity on block-based storage, allowing recovery of production applications with minimal data exposure.

Application owners can tune recovery point objectives based on criticality of data and perform faster recovery through self-service capabilities. Copies of production data can be used for development, testing, decision support tools, reporting, and backup acceleration:

- SnapView
- SnapSure™
- RecoverPoint/SE continuous data protection (CDP)

Replication management tools

Application Protection Suite automates the creation of application-consistent restartable database copies, so that SQL Server database can be recovered to defined service levels. For recoverable database copies, SQL Server transaction logs need to be backed up separately:

- For **VNX series** to include Replication Manager/AppSync and Data Protection Advisor for Replication Analysis.
- For **VNXe series** to include Replication Manager/AppSync.
- For **VMAX series** to include Replication Manager/AppSync.

To achieve point-in-time recoverability for replication manager or AppSync, the full backups taken with the snapshot or clones need to be combined with SQL Server log backups.

Restoring a Database to a Point Within a Backup, available from the [Microsoft MSDN Library](#), provides more information about point-in-time recovery.

Multi-site disaster recovery

Considerations

The most important requirements for implementing a multi-site disaster recovery solution usually in Service Level Agreement (SLA) are:

- **Recovery time objective (RTO):** How long can the end user of the SQL Server tolerate the interruption of service.
- **Recovery point objective (RPO):** How much data loss can be tolerable.
- **Cost:** Cost of the solution to make that service-level agreement (SLA) feasible.

With a **synchronized replication solution**, data is only acknowledged when the remote site data is committed:

- **Advantage:** Zero data loss (0 RPO) at any time.
- **Disadvantage:** This can potentially slow down the production environment with slow link over a long distance. Synchronous replication over distances greater than 200 km may not be feasible.

An asynchronous replication solution will not have the distance limitation because the data will be committed before the remote site sends acknowledgement:

- **Advantage:** No limitation for replication distance.
- **Disadvantage:** It can have potential data loss.

The amount of data/volumes protected is another consideration of the multi-site protection design:

- All the data is replicated.
 - It can be set up on the remote site to automatically start when failover is needed, providing instant RTO.
 - More data transferring over the network might degrade production performance.
- Choose only user database and log files to replicate.
 - Less data transferring over the network means better production performance and less data loss.
 - Less data in the consistency group can potentially prolong the recovery procedure and time (longer RTO).

For the highest level of RTO and RPO, choose the synchronized solution, possibly with geographically dispersed clustering product, this will provide zero data loss solutions with extremely small RTO, with most processes automated (VMAX SRDF/CE). This needs invest on fast links between the sites.

For a higher level of RTO and RPO, VPLEX provides similar results with heterogeneous arrays at the remote site.

RecoverPoint, Replicator, and the rest of EMC multi-site replication technologies can all provides very good RTO and RPO with minimum user intervention to bring the remote site up when disaster recovery is needed.

Multi-site replication technologies

To extend the functionality of single-site Windows failover cluster configurations and provide additional multisite protection, EMC provides the following solutions:

- VMAX SRDF/Cluster Enabler for MSCS geographically dispersed clustering product.
- RecoverPoint provides both synchronous and asynchronous remote protection. Asynchronous protection can replicate data across any distance.
- VPLEX provides access to a single copy of data at different geographical locations concurrently, enabling a transparent migration of running virtual machines between data centers.
- VNX Remote Protection Suite provides protection through Replicator, MirrorView™, and RecoverPoint/SE continuous remote replication (CRR).
- VNXe Replicator provides remote protection for iSCSI and NAS

VNX Remote Protection Suite

VNX Remote Protection Suite delivers unified block and file replication, providing disaster recovery for both NAS and SAN environments. It delivers disaster recovery protection for any host and application without compromise—with immediate DVR-like rollback to a point in time. Capabilities include compression and deduplication for WAN bandwidth reduction, application-specific recovery point objectives, and replication options for one-to-many configurations.

- This suite for VNX series includes Replicator, MirrorView/A, MirrorView/S, and RecoverPoint/SE continuous remote replication (CRR).
- This suite for VNXe series includes replicator (iSCSI and NAS).

MirrorView replicates SQL Server database LUNs to remote locations for disaster recovery. MirrorView replication is transparent to the host. If the production host or the production storage system fails, the remote replication facilities failover to the secondary mirror image.

MirrorView software offers two complementary mirroring products:

- MirrorView/S can synchronously mirror data images of production host LUNs to secondary storage at a remote site in real time. This offers zero data loss if there is a failure at the production site.
- MirrorView/A offers long-distance replication based on a periodic incremental update model. It periodically updates the remote copy of the data with all the changes that occurred on the local copy since the last update. This can result in data loss if there is a failure at the production site.

MirrorView works well in small- to medium-sized SQL Server environments. *EMC Business Continuity for Microsoft SQL Server 2008 Enabled by EMC CLARiiON and EMC MirrorView /A White Paper* on the EMC website provides more information about MirrorView.

Tools to restart automation

Microsoft Failover Clustering instance (FCI) provides restart automation by bringing up the secondary site automatically if the primary site server goes down.

VMAX SRDF/CE for MSCS provides a geographically dispersed, clustering solution with a high level of automation that provides zero data loss and extremely small RTO for SQL Server disaster recovery on a different site at the instance level.

Microsoft AlwaysOn Availability Groups provide disaster recovery for servers on multiple sites with a choice of automatic failover at the database level if the primary database goes down.

Virtualized instances automation tools

The vCenter Site Recovery Manager (SRM) provides automated disaster recovery for fast and efficient recovery of critical applications, such as SQL Server, by simplifying recovery and eliminating human error from the process. The sample design and reference architecture in [vCenter SRM solution protection](#) provides SRM configuration details.

Disaster recovery options for SQL Server 2012

EMC offers various DR options for SQL Server 2012. Table 15 describes some frequently used options. Each option has advantages and disadvantages. The option that is best for an environment is determined by its specific DR requirements.

Table 15. EMC disaster recovery offerings for SQL Server

Offering	Replication method	Description
AlwaysOn Availability Groups	SQL Server native continuous replication	Built-in to SQL Server 2012 for high availability and disaster recovery
Database portability	EMC RecoverPoint	Only SQL Server data is replicated. Requires DNS changes when failover to the secondary replica
	EMC VPLEX	
Server/site move	EMC RecoverPoint	Both OS and SQL Server data are replicated and failover includes server start, IP change, and DNS update
	EMC VPLEX	
	EMC VMAX SRDF/Cluster Enabler	
	EMC Replicator	

Additional backup recommendations

Follow these additional recommendations for SQL Server backups to reduce performance degradation:

- With medium to high workloads, do not take backups directly from the production SQL Server. Instead, mount a point-in-time snapshot or clone on a different server and take the backups from that server or with a secondary copy in an AlwaysOn Availability Group.
- Schedule backups to occur during off hours, whenever possible.

AlwaysOn for HA/DR

The AlwaysOn Availability Group is a Microsoft SQL Server native continuous replication solution that is built into SQL Server 2012 for high availability and disaster recovery. Implementation for AlwaysOn Availability Groups fits seamlessly with EMC flash-based solutions, such as FAST VP, FAST Cache, flash XtremSW Cache, and XtremSF.

AlwaysOn with FAST Suite

When AlwaysOn Availability Groups are configured as part of the HA and DR plan for the SQL Server database, the following are best practices for the secondary copy design of AlwaysOn Availability Groups:

- If the secondary copy does not need to support heavy reporting workload, it can reside on lower tier storage if it can sufficiently support the required workload when a failover occurs.
- When FAST VP or FAST Cache is enabled for the secondary copy, the read-only workload can be improved. If the secondary copy does not share the pool with the primary, there will be no impact on the primary site. This also allows for the highest level of performance when a failover requires the secondary copy.

- The *EMC Mission Critical Infrastructure for Microsoft SQL Server 2012 White Paper* provides a sample design and reference architecture and other implementation details.

AlwaysOn with flash XtremSW Cache/XtremSF

When server-side flash is implemented, the performance boost for SQL Server workloads allows for much higher performance with extremely low latency. The best practices and design considerations for XtremSW Cache and XtremSF are as follows:

- XtremSW Cache usually only needs to be configured on the primary site unless the secondary site requires a performance boost in read-only workload (only when the workload is random) or higher performance is required for failovers to the secondary site.
- If the secondary copies in AlwaysOn Availability Groups also have XtremSF with a high-speed network and short distance, synchronized copy can be implemented with minimum impact on the production server and minimum data loss when a failover occurs.
- Multiple secondary copies of AlwaysOn Availability Groups have less impact on the production server when XtremSF is used for both primary and secondary sites with a fast network link.
- XtremSW Cache can be used in conjunction with EMC storage FAST Suite for further boosting SQL Server performance.

The *EMC XtremSF Acceleration for Microsoft SQL Server 2012 White Paper* provides more details about the XtremSF design and implementation.

The *EMC Infrastructure for High Performance Microsoft and Oracle Database System White Paper* provides more details about the XtremSW Cache design and implementation.

Conclusion

Summary

This document highlights the key decision points in planning a Microsoft SQL Server deployment with EMC storage systems. Multiple configuration options are available to suit most environments. EMC storage and data management products are designed for flexible management of SQL Server environments to best meet your business needs.

Best practices for designing SQL Server storage are constantly evolving. With storage technology rapidly improving, traditional best practices may not apply to all configurations. This document presents the current best practices recommended by EMC for deploying SQL Server with the EMC VNX family of unified storage or EMC Symmetrix VMAX storage series. Following these guidelines can greatly assist you in achieving an efficient, high-performance, and highly available SQL Server environment that meets your requirements.

This paper presents the following concepts, principles, and formulas to help you:

- Understand the I/O and bandwidth characteristics of SQL Server.
- Apply best practices for SQL Server and the VNX or VMAX storage series.
- Utilize a SQL Server storage building block.
- Calculate the storage I/O, capacity, and bandwidth requirements.
- Validate your overall storage design.
- Become familiar with the various data protection options for SQL Server.

Additional information

Consult your local EMC SQL Server expert for additional guidance on deploying Microsoft SQL Server with the EMC VNX family or EMC Symmetrix VMAX storage series.

Appendix A: EMC Data Protection Advisor for Replication Analysis

Overview

EMC Data Protection Advisor for Replication Analysis (DPA/RA) automates the collection of data from applications, hosts, and arrays; constantly monitors for exposures; and sends alerts for potential missed SLAs and gaps in the protection objectives.

Data Collection and Discovery wizards

Monitoring devices and applications are automated with the **Data Collection** wizard and **Discovery** wizard, which configures DPA/RA with a series of questions about the device or application to monitor. After defining a device or application with the wizards, one or more nodes are automatically added to the **Configuration** view and data monitoring for when the collector starts.

Data discovery and collection

In this example, VNX arrays are remotely discovered and monitored from the collector running on the DPA/RA server. The VNX system is monitored for recoverability and analysis reporting.

Discovering storage arrays from DPA/RA requires EMC Solutions Enabler to be installed. To install Solutions Enabler, use the following steps:

1. Install Solutions Enabler on the DPA/RA server.
2. Create a text file with the following information, one line per array (in this example, the file name is Clar.txt):

```
<SPA IP> <SPB IP> <Username> <Password>
```

3. To register the VNX, run the following command on the DPA/RA server:

```
C:\Users\administrator>symcfg disc -clar -f c:\Clar.txt
This operation may take up to a few minutes. Please be patient...
Discovering Clarion at SpA: 172.30.238.23 and SpB: 172.30.238.23 ... Done
```

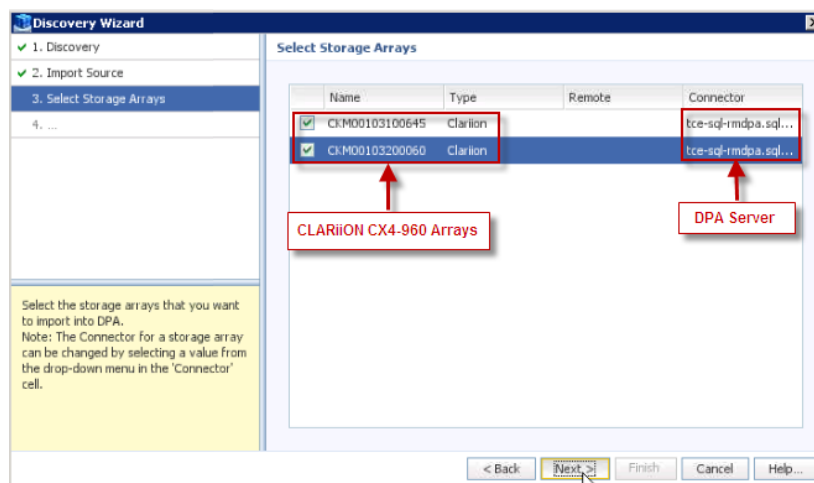
4. To verify the VNX was added successfully, run the following command:

```
C:\Users\administrator>symcfg list -clar
                C L A R I O N
ClarID          Model      Firmware Version      Num Disks  Num Phys Devices  Num Clar Devices
CKM00103100645 CX4_960 4.30.0.5.509         90          0                54
CKM00103200060 CX4_960 4.30.0.5.509        120          1                86
```

Discovering storage arrays

To discover storage arrays with DPA/RA, follow these steps:

1. From the DPA/RA toolbar, select **Tools**, then choose **Discovery Wizard**.
2. Select **Storage Arrays**, proceed to the **Import Source** panel, click **Next**. The Discovery wizard displays a list of all of the storage arrays. Select the storage arrays to import and click **Next**.



3. Select a **Schedule** for the recoverability data gathering request and click **Finish**.

Configuring Data Protection Advisor for Microsoft SQL Server monitoring

To configure DPA/RA for SQL Server monitoring, use the following steps:

1. From the DPA/RA toolbar, select **Tools** then choose **Data Collection Wizard**.
2. Click **Host** and click **Next**. The **Host Details** panel appears.

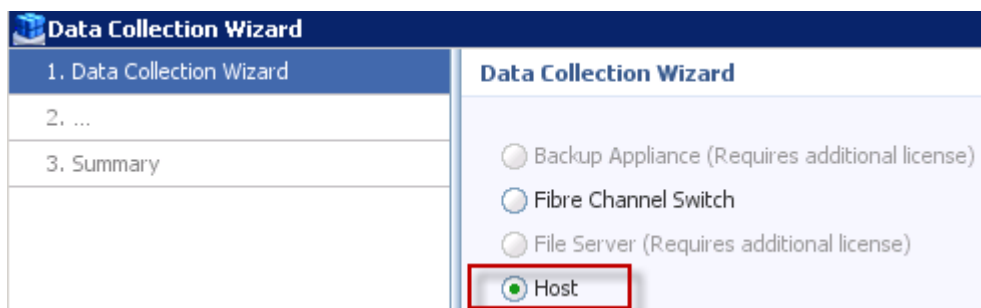


Figure 9. Data Collection Wizard

3. Enter the name, description and OS type of the host.
4. Under Collector Location (Figure 10), choose Yes for “Is there or will there be a Collector installed on the Host?”

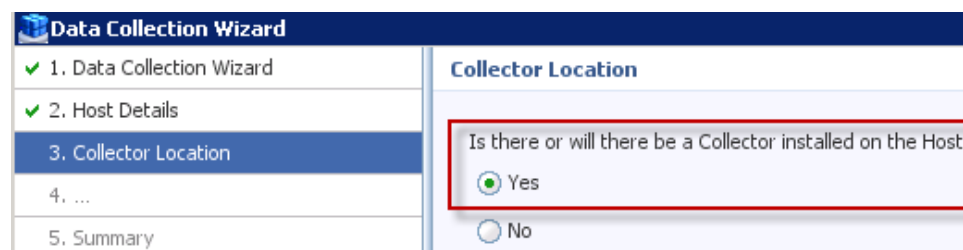


Figure 10. Collector Location

5. To gather CPU performance and memory utilization data from the OS, under **Data Gathering**:

- a. For **Do you want to gather system information?**, choose **Yes**.
- b. For **Do you want to monitor applications on this host?**, choose **Yes**.
- c. Check **Microsoft SQL Server**, as shown in Figure 11.

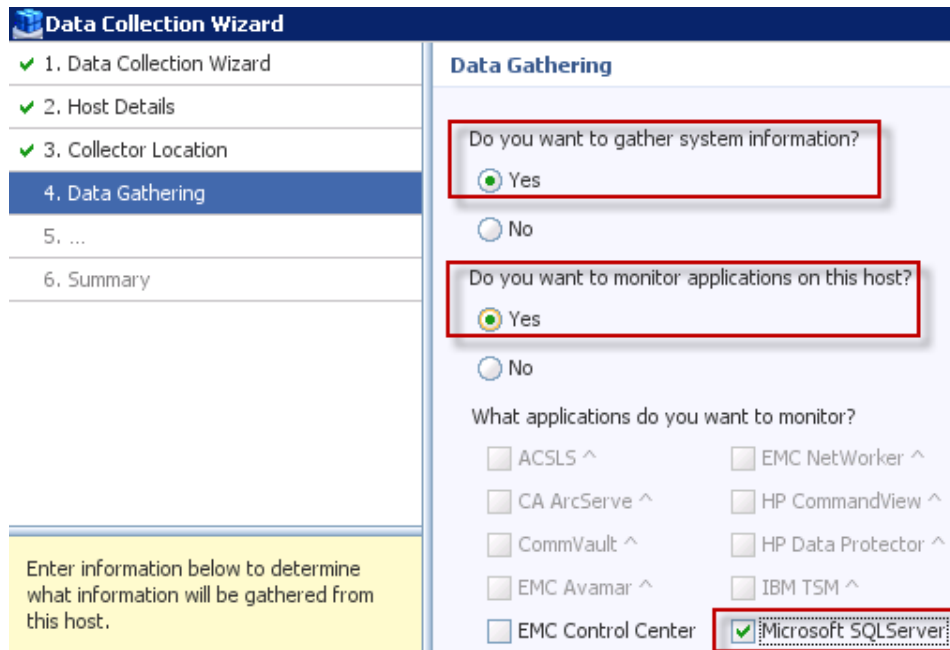


Figure 11. Data Gathering

6. To add a SQL Server instance, in the **Data Gathering** pane, click **Add**. The **Add SQL Server Instance** dialog box appears. Enter the applicable SQL server credentials and close the dialog box.

Figure 12 shows the SQL Server was successfully added to the DPA/RA configuration.

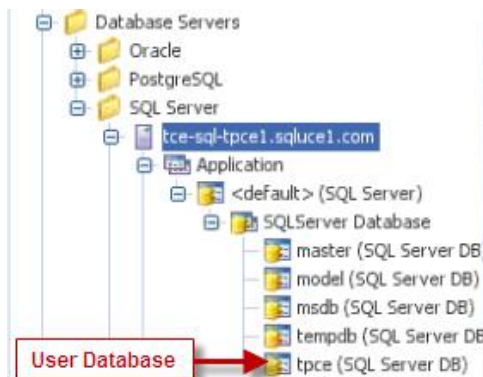


Figure 12. DPA/RA database server view

Displaying and reporting gaps and exposures

DPA/RA provides an intuitive graphical map of the relationship between the host and storage. DPA/RA presents the recoverability gaps and exposures by using reports and views to resolve issues. DPA/RA can monitor numerous replication error conditions.

Figure 13 shows the configuration for setting up a scheduled report, while Figure 14 shows the exposure details for SQL Server. This maps the storage to RecoverPoint and then into the SQL Server virtual machine residing on ESX cluster.

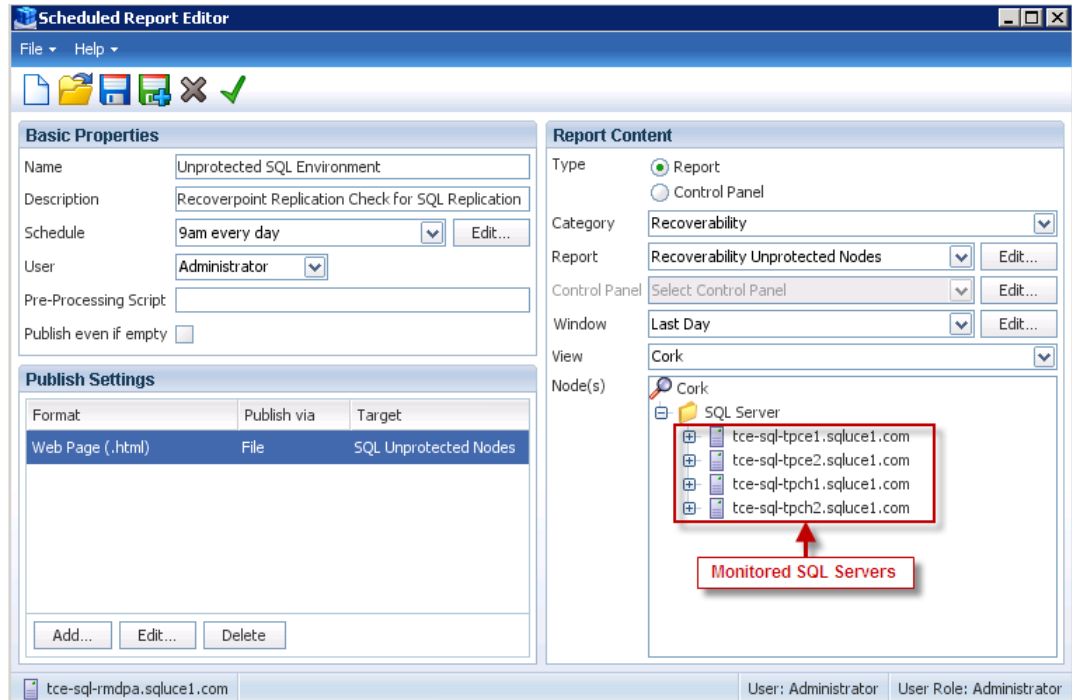


Figure 13. Scheduled Report Editor showing monitored SQL Server systems

Recoverability Unprotected Nodes - : (20/05/11 15:10)

Client	Managed Object
tce-sql-tpce1.sqluce1.com	Default (MSSQL): dpa test db

Figure 14. Exposure details for a SQL Server

From this, DPA/RA detects that the configuration is missing an application volume, the replica is incomplete, and the application might not be recoverable. You can then check the configuration and correct it accordingly.

Appendix B: Tools for SQL Server performance monitoring, tuning, and sizing

Overview

SQL Server can be monitored in many levels, from top down, including: Application that uses SQL Server for transactions, SQL Server database, Windows host that hosts the SQL Server, Hypervisor (if virtualized), and the storage layer where all the data resides.

EMC recommends gathering performance data while production workloads are running. In some cases, the workload characteristics change periodically (for example, a system might do OLTP during the day, ETL and reporting in the evening, and backup at night). In this case, you should capture all phases of the day, so you can size a system to address all phases of production.

When running a Windows failover cluster, EMC recommends gathering data from all nodes simultaneously, so that performance data is gathered even during a cluster failover.

EMC also recommends collecting and analyzing counters that indicate memory and CPU pressure on the physical or virtual machine, as these factors can affect the storage performance. For example, adding memory to a server can significantly reduce storage I/O or removing a CPU bottleneck can significantly increase storage I/O. Table 16 lists the tools for each level of use.

Table 16. Tools used for SQL Server performance monitoring, tuning, and sizing

Level	Tool	Source/Links	Description
Application	DBclassify	EMC (http://www.emc.com/domains/zettapoint/index.htm)	Constantly monitors data, learns its patterns and past behavior, and then classifies and moves it according to business priorities.
	Perfcollect	EMC (http://emc.ms/Perfcollect)	Automates SQL server related performance data collection. Mainly used for storage and virtual environment sizing.
	EMC workload performance assessment	EMC (https://emc.mitrend.com)	Also known as “Mitrend.” Automated online workload performance assessment tool, which correlates and displays key performance information related to sizing.
	PAL	Performance Analyzer of Logs—Open source (http://www.codeplex.com/PAL)	Useful for troubleshooting performance issues.

Level		Tool	Source/Links	Description
SQL Server database		VSPEX SQL sizing tool	EMC (http://express.salire.com/go/emc)	Can be used to determine the recommended VSPEX Proven Infrastructure for virtualized SQL Server based on the user requirements.
		T-SQL	Microsoft (comes with SQL server installation)	Provides Transact-SQL system stored procedures to create traces on an instance of the SQL Server Database Engine.
		SQL Server profiler	Microsoft (SQL Server analysis Services)	Provides SQL Trace capture and replay in a graphic user interface.
		SQL Database Tuning Advisor(DTA)	Microsoft (SQL Server analysis Services)	DTA provides SQL Server tuning suggestion such as indexing and partitioning.
		Dynamic Management Views (DMVs)	Microsoft (SQL Server analysis Services)	Dynamic Management Views are query structures that expose information about local server operations and server health.
Windows host		Perfmon	Windows performance monitor (comes with windows server installation)	Perfmon can track the performance characteristics of SQL Server workloads.
Hypervisor	VMware	vSphere Client GUI interface	vSphere client GUI	Primary tool to track performance and configure data for one or more ESX/ESXi hosts.
		Resxtp/Esxtp	ESX/ESXi	Provides a performance matrix, but requires root access.
	Hyper-V	Perfmon	Windows performance monitor	Provides a performance matrix for Hyper-V and virtual machines.
Storage/Server Cache		Unisphere Analyzer	Included with EMC storage systems	Provides performance monitoring for EMC storage systems.
		XtremSW Cache performance predictor tool	https://support.emc.com/search/?product_id=25208&text=predictor	Provides a performance predictor tool for EMC XtremSW Cache to assess and evaluate the SQL Server environment for the XtremSW Cache.

Level	Tool	Source/Links	Description
	EMC Storage Configuration Advisor	Available through EMC pre- and post-sales	Assists in defining tiering policies for an existing environment; Tier Advisor monitors I/O and recommends tiering policy settings.

Application-level tools

EMC DBclassify

EMC DBclassify™ is a database optimization solution that reduces the total cost of ownership of database storage, while enhancing the performance of business applications. DBclassify constantly monitors data, learns the patterns and past behavior, and then classifies and moves the data according to business priorities. DBclassify is ideal for IT organizations facing management, performance, and budgetary challenges associated with increasingly complex databases.

DBclassify analyzes and differentiates structured data in order to provide full visibility into the actual use of the database. Through a comprehensive analysis process, DBclassify automatically tracks and ranks every database object (tables, indexes, and partitions), based on frequency of access and I/O wait information. Additionally, DBclassify associates every database object with users and applications based on actual usage.

Using a unique ranking formula developed by DBclassify provides an optimal storage tiering solution for databases across the enterprise, offering tiering recommendation for the object, table space, or file level, as well as functioning as the policy engine for EMC FAST technology.

In SQL Server environments, data is collected with a remote database connection instead of a monitored server-based agent. The collector process runs on the DBclassify repository server and pulls information from the monitored database. Figure 15 shows the DBclassify architecture for SQL Server.

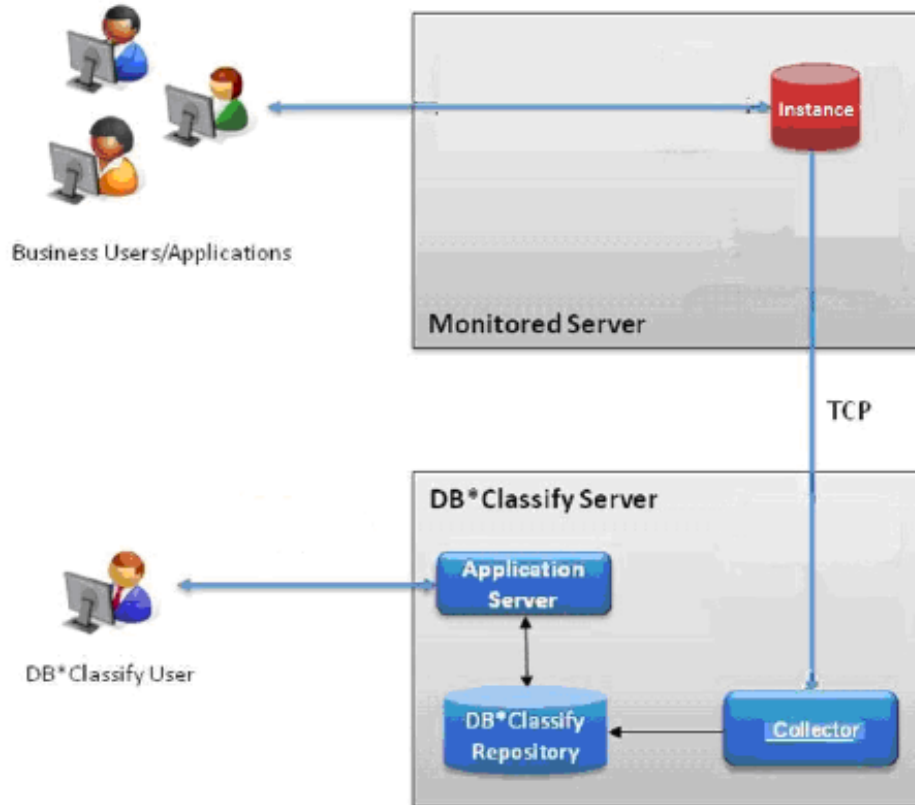


Figure 15. DBclassify SQL Server architecture

Best practices with DBclassify are as follow:

- Deploy DBclassify before a major upgrade/migration to benchmark current I/O profile.
- Capture peak workloads for detailed I/O analysis.
- Generate FAST VP policy based on varying levels of threshold percentages and tiered storage capacity.
- Create business filters that reflect the customers workloads and processes.
- Analyze usage data based on business filters.
- Classify usage data into hot, warm, or cold categories.
- Profile and benchmark I/O activity across all tiers for all databases.
- Group databases into classification based on importance.
- Profile and benchmark storage tiering capacity by database classification.

You can find more information about DBclassify on the [DBclassify website](#):

This tool is available through your reseller or EMC presales systems engineer.

Perfcollect

EMC provides an automated tool set called **Perfcollect**, which is available at no charge. Perfcollect automates the collection of SQL Server, storage, memory, and CPU counters, along with other configuration information that assists in storage and virtual environment sizing. The tool can be run on any server running Windows 2003 or later.

You can manually analyze the performance monitor information with the Windows Performance monitor.

EMC Workload Performance Assessment Tool

The EMC Workload Performance Assessment tool is available to EMC partners. This automated online tool correlates and displays key performance information related to sizing. This tool is available from your reseller or EMC presales systems engineer.

PAL

You can use the open source Performance Analyzer of Logs (PAL) tool to troubleshoot performance issues (as opposed to sizing for migrations). You can use the data collected with Perfcollect with the PAL tool, which can be downloaded from the [CodePlex website](#).

SQL Server database-level tools

VSPEX SQL Server sizing tool

The VSPEX SQL Server sizing tool is available to size VSPEX Proven Infrastructure for virtualized SQL Server 2012 based on user requirements. You can also use the sizing estimate for other virtualized environments on EMC VNX storage.

Figure 16 through Figure 20 provide sample outputs for sizing a 500 GB OLTP database with a maximum of 4,000 IOPS and 10 percent annual growth on a VNX5400 system with reference virtual machines (RVMs).

VSPEX Private Cloud – VMware up to 300 RVMs for next generation VNX					
Recommended array	VNX5400				
Max RVMs for array	300				
Storage RVMs Required	1				
Storage RVMs remaining in this configuration	10				
Additional disk number required for application data	16				
Compute RVM Required	16				
% VSPEX Model Utilization	10%				

Application Name	Total vCPU of RVM	Total Memory of RVM	Total OS Volume Cap of RVM	Total OS Volume IOPS of RVM	Total RVM
SQL Server 2012	8	16	1	1	16

Figure 16. VSPEX configuration

Infrastructure	Disk Type	Disk Size (GB)	Number of Drives
VSPEX Private Cloud Pool			
SAS Tier	15K SAS	600	5
Flash Tier	FAST VP SSD	100	2
Hot Spares			
	SSD	200	1
	15K SAS	600	1

Application Data	Total Drives
SQL Data Storage Pool	16
TOTAL	16

Figure 17. VSPEX disk requirements

Instance #1	Temp DB Size	Avg Annual Growth Rate	Include FAST Cache?
SQL Server Instance 1	0	10%	Yes

Database Profile	Maximum Database Size (GB)	Maximum Database Performance (IOPS)	Transactions/sec at peak load (optional)
DB1	500	4,000	0

Figure 18. SQL Server 2012 inputs (Instance No.1)

Role	# of VMs	vCPU of RVM	Memory of RVM	OS Volume Cap of RVM	OS Volume IOPS of RVM	Total RVM
SQL Server Instance 1	1	8 vCPUs (8 RVM)	32 GB (16 RVM)	100 GB (1 RVM)	25 IOPS (1 RVM)	16

Figure 19. SQL Server 2012 resource requirements

Pool Name	Disk Type	Disk Size (GB)	# of Disks	RAID
SQL Server Instance 1				
SQL Database Pool (with FAST Cache Enabled) for SQL Server Instance 1	15K SAS	300	40	RAID 5 (4+1)
SQL TempDB and Log Pool for SQL Server Instance 1	15K SAS	300	8	RAID 1 (1+1)

Figure 20. SQL Server disk requirements

Transact-SQL

Microsoft SQL Server provides Transact-SQL system stored-procedures to create traces on an instance of the SQL Server database. These system-stored procedures can be used within your own applications to create traces manually, instead of using SQL Server Profiler. This allows you to write custom applications specific to the needs of your enterprise. Table 17 lists the stored procedures for tracing an instance of the SQL Server Database Engine.

Table 17. Transact-SQL stored procedures for SQL server trace

Stored procedure	Task performed
fn_trace_geteventinfo (Transact-SQL)	Returns information about events included in a trace.
fn_trace_getinfo (Transact-SQL)	Returns information about a specified trace or all existing traces.

Stored procedure	Task performed
sp_trace_create (Transact-SQL)	Creates a trace definition. The new trace will be in a stopped state.
sp_trace_generateevent (Transact-SQL)	Creates a user-defined event.
sp_trace_setevent (Transact-SQL)	Adds an event class or data column to a trace or removes one from it.
sp_trace_setstatus (Transact-SQL)	Starts, stops, or closes a trace.
fn_trace_getfilterinfo (Transact-SQL)	Returns information about filters applied to a trace.
sp_trace_setfilter (Transact-SQL)	Applies a new or modified filter to a trace.

[Microsoft SQL trace](#) provides more details.

SQL Server Profiler

SQL Server Profiler is a rich interface to create and manage traces and analyze and replay trace results. The events are saved in a trace file that can later be analyzed or used to replay a specific series of steps when trying to diagnose a problem.

SQL Server Profiler can be used for the following tasks:

- Stepping through problem queries to find the cause.
- Finding and diagnosing slow-running queries.
- Capturing the series of Transact-SQL statements that lead to a problem.
- Diagnosing problem on a test server with replaying the saved trace.
- Monitoring SQL Server performance to tune the workloads.
- Correlating performance counters to diagnose problems.

[Microsoft SQL Server Profiler](#) provides more details.

SQL Server Database Engine Tuning Advisor

The Microsoft Database Engine Tuning Advisor (DTA) analyzes databases and makes recommendations to optimize query performance. The DTA can be used to select and create an optimal set of indexes, indexed views, or table partitions without having an expert understanding of the database structure or the internals of SQL Server.

Using the DTA, you can perform the following tasks:

- Troubleshoot the performance of a specific problem query
- Tune a large set of queries across one or more databases
- Perform an exploratory what-if analysis of potential physical design changes
- Manage storage space

For more information about tuning the physical database design for database workloads, see [Database Engine Tuning Advisor](#).

SQL Server Dynamic Management Views

SQL Server Dynamic Management Views (DMVs) are query structures that expose information about local server operations and server health. The query structure includes interface to schema row sets that return metadata and monitor information about an Analysis Services instance.

Windows host-level tool

Windows Performance monitor (Perfmon)

Windows Perfmon tracks the performance characteristics of workloads running on physical and virtual machines, as well as Hyper-V. It can be used in real-time to view current performance, and can also be configured to log performance data to a file for later viewing and processing. Because it is collected from the same operating environment as the application, the performance monitor most closely reflects performance as viewed by the application.

Table 18 lists the most useful counters for evaluating the activity and performance of storage in block (Fibre Channel, iSCSI, and SAS) environments, which you can view from either the **PhysicalDisk** or **LogicalDisk** counter sets.

Table 18. Useful counters for evaluating storage performance in SAN environments

Counter	Measured feature
Avg disk sec/transfer	Overall Storage Latency
Avg disk sec/read	Read Latency
Avg disk sec/write	Write Latency
Avg disk bytes/transfer	I/O Size
Disk bytes/sec	Throughput
Disk reads/sec	Read I/O/sec
Disk writes/sec	Write I/O/sec

SQL Server also supports the SMB protocol for databases and transaction logs, which requires the **SMB Client Shares** counter set, as shown in Table 19.

Table 19. Useful counters for evaluating storage performance in NAS environments

Counter	Measured feature
Average data request per second	Overall storage latency
Average read per second	Read latency
Average write per second	Write latency
Average data bytes per request	I/O size
Data requests per second	Throughput
Read requests per second	Read I/O/sec
Write requests per second	Write I/O/sec

Microsoft has defined generally acceptable storage latencies. Table 20 lists guidelines for further investigation; certain applications may benefit from lower latencies or tolerate higher latencies.

Table 20. Guideline latencies for data storage in SQL Server OLTP environments

Workload	Average latency	Peak latency
Database	<20ms read	<50ms read
Transaction logs	<10ms write	<50ms write
System/Page partition	<10ms read/write	<10ms read/write

The counter Page Life Expectancy (PLE) is a good indicator for memory pressure (minimum of 300 ms).

More information on performance monitor counters related to SQL Server and storage can be found here: [SQL Server performance monitoring](#).

Hypervisor-level tools

Key metrics to monitor ESX

In the VMware environment, the following are two ways to monitor ESX/ESXi performance:

- vSphere Client GUI interface:
 - Primary tool for observing performance and configuring data for one or more ESX/ESXi hosts
 - Does not require high levels of privilege to access the data

- Resxtop/Esxtop:
 - Gives access to detailed performance data for a ESX/ESXi host
 - Provides fast access to a large number of performance metrics
 - Requires root-level access
 - Runs in interactive, batch, or replay mode

Table 21 shows the key metrics to monitor in the ESXi environment for both the hosts and virtual machines.

Table 21. Key metrics to monitor for ESXi hosts and virtual machines

Resource	Metric	Description
CPU	%USED	CPU used over the collection interval (%)
	%RDY	CPU time spent in the ready state (virtual machine only)
	%SYS	Percentage of the time spent in the ESXServer virtual machine Kernel
Memory	Swapin, Swapout	Memory ESX host swaps in/out from/to disk (per virtual machine, or cumulative over host)
	MCTLSZ (MB)	Amount of memory reclaimed from resource pool by way of ballooning
Disk	Reads/s, Writes/s	Reads and writes issued in the collection interval
	DAVG/cmd	Average latency (ms) of the device (LUN)
	KAVG/cmd	Average latency (ms) in the virtual machine kernel, also known as queuing time
	GAVG/cmd	Average latency (ms) in the guest. GAVG = DAVG +KAVG
Network	MbRX/s, MbTX/s	Amount of data transmitted per second
	PKTRX/s, PKTTX/s	Packets transmitted per second
	%DRPRX, %DRPTX	Dropped packets per second

Key metrics to monitor Hyper-V

Hyper-V performance can be monitored with Perfmon. Table 22 shows the key metrics to monitor in the ESXi environment for both the hosts and virtual machines.

Table 22. Key metrics to monitor for Hyper-V hosts and virtual machines

Resource	Metric	Description
CPU	\Processor(*)\% Processor Time \Hyper-V Hypervisor Logical Processor(_Total)\% Total Run Time	CPU time busy Less than 60% consumed = Healthy 60% - 89% consumed = Monitor or Caution 90% - 100% consumed = Critical, performance will be adversely affected
Memory	\Memory\Available Mbytes	The amount of physical memory available to the Hyper-V host 50% of free memory available or more = Healthy 25% of free memory available = Monitor 10% of free memory available = Warning Less than 5% of free memory available = Critical, performance will be adversely affected
	\Memory\Pages/sec	Measures the rate at which pages are read from or written to disk to resolve hard page faults. Less than 500 = Healthy 500 - 1000 = Monitor or Caution Greater than 1000 = Critical, performance will be adversely affected
Disk	\Logical Disk(*)\Avg. sec/Read, \Logical Disk(*)\Avg. sec/Write	Read and write latency. 1ms to 15ms = Healthy 15ms to 25ms = Warning or Monitor 26ms or greater = Critical, performance will be adversely affected

Resource	Metric	Description
Network	\Network Interface(*)\Bytes Total/sec	The percentage of network utilization Less than 40% of the interface consumed = Healthy 41%-64% of the interface consumed = Monitor or Caution 65-100% of the interface consumed = Critical, performance will be adversely affected
	Network Interface(*)\Output Queue Length	The output queue length measures the number of threads waiting on the network adapter 0 = Healthy 1-2 = Monitor or Caution Greater than 2 = Critical, performance will be adversely affected

For detailed information about hyper-V implementation, refer to the *Virtualizing SQL Server 2008 using EMC VNX Series and Microsoft SQL Server 2008 R2 Hyper-V White Paper*.

Storage/Server cache-level tools

Unisphere Analyzer

Unisphere Analyzer is a performance monitoring tool for EMC storage arrays. This tool not only shows the storage specific performance, but also collects and shows some server and virtual machine level performance information.

For more details, see the *EMC Unisphere Unified Storage Management Solution White Paper*.

XtremSW Cache Performance Predictor

EMC XtremSW Cache Performance Predictor is a tool that can be used to estimate the benefits of implementing XtremSW Cache in a specific environment.

The tool is run in two steps:

1. Data collection on the host side using common trace collection tools.
2. Trace analysis on a host or on any laptop that meets the system requirements. The tool simulates the XtremSW Cache way of operation and generates PDF output file.

No card or software purchase is required to run this free tool that can run on all XtremSW Cache supported operating systems (Windows and Linux). It creates a set of charts and graphics to show whether XtremSW Cache can benefit the environment, and then estimates performance improvement based on:

- Observed host response time
- Capacity used by host
- Skew level

Figure 21 shows the performance collection and the cache configuration from a sample PDF output of the tool.

XtremSW Cache Performance Predictor

The EMC XtremSW Cache Performance Predictor simulation shows that your environment will significantly benefit from implementing a 700GB XtremSW Cache caching solution. The resulting latency improvement of the simulated workload is estimated at 79.76%.

This estimate is for illustration purposes only, and is not meant to predict the actual behavior of XtremSW Cache in this environment.

Simulation Properties:

Server name: WIN2008R2SP1

LUN: harddisk1

Trace file: xperf_trace2.csv

Sampled window: 22/12/2012, 18:40:46 - 18:41:24

Total number of I/O requests: 140,542

Percentage of total array I/Os (in the sampled window): 99.73%

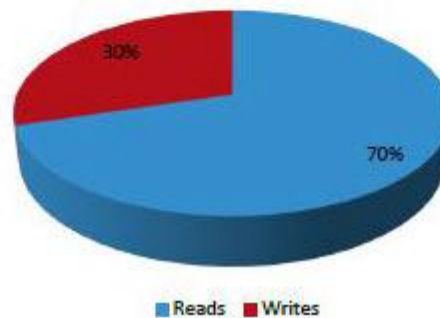
Simulated cache properties:

Cache size: 700GB

Cache page size: 8KB

Max IO: 64KB

Workload Mix



XtremSW Cache will provide greater benefits as the read/write ratio increase.

Figure 21. XtremSW Cache Performance Predictor sample output: collection of performance data

Figure 22 shows the output of the tool the I/O distribution of the disk. This information can be used to set the page size and maximum I/O size of actual XtremSW Cache for better performance if needed (the default page size is 8 KB and the maximum I/O size is 64 KB).

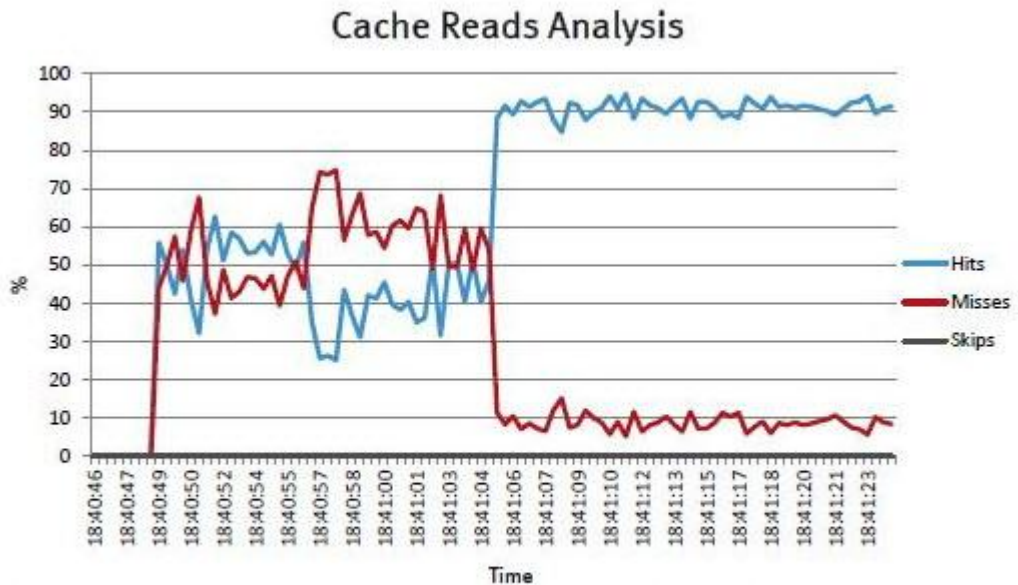
Server name: WIN2008R2SP1
LUN: harddisk1



XtremSW Cache can be configured to cache I/O sizes of up to 128KB.

Figure 22. XtremSW Cache Performance Predictor sample output: I/O size distribution

Figure 23 shows the cache read analysis. If the tool gives a very high cache hit rate, then this device under load is a good candidate for XtremSW Cache acceleration.



Higher read hits means that a larger part of the application working set is promoted to XtremSW Cache, resulting in higher performance.

Figure 23. XtremSW Cache Performance Predictor sample output: prediction of cache hit

Figure 24 shows an estimate of the performance improvement that the disk can get from the XtremSW Cache acceleration. This is a simulated result and serves a good reference for how the application benefits from the XtremSW Cache acceleration.

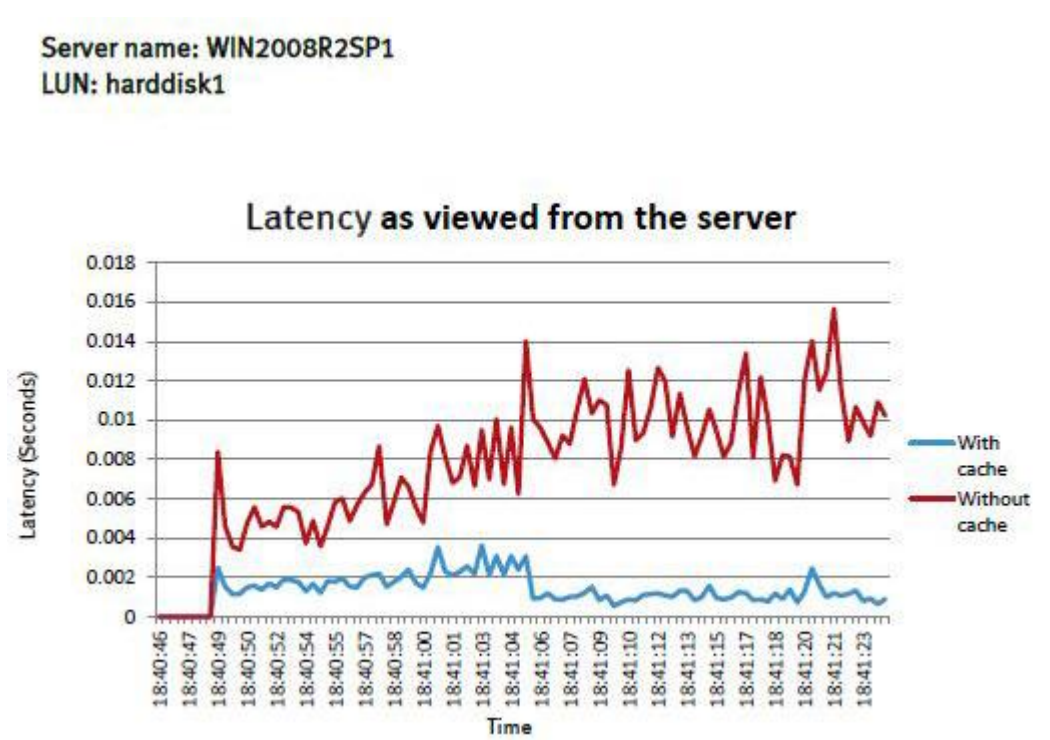


Figure 24. XtremSW Cache Performance Predictor sample output: disk latency prediction

XtremSW Cache Performance Predictor should be used as a planning tool when designing XtremSW Cache for best performance.

EMC Storage Configuration Advisor

EMC Storage Configuration Advisor is a new EMC Storage Resource Management (SRM) Suite that delivers one of the industry's most comprehensive application-to-storage management, for common insight into service level management, priorities and tasks as customers build their cloud infrastructure.

The SRM Suite combines EMC ProSphere, EMC Storage Configuration Advisor, and recently acquired EMC Watch4net into a single, easily consumable monitoring and reporting package. This package offers performance, capacity, and configuration management at scale for EMC and selects third party storage arrays for both file and block.

Storage Configuration Advisor provides the following advantages:

- **Best-practice policy management**

Storage Configuration Advisor supplies built-in templates around common industry best practices to enable you to define and modify policies that align with operational requirements.

- **Change tracking**

Storage Configuration Advisor gets alerts when problems occur that put service levels at risk through continuous change tracking and configuration validation.

- **Configuration analysis**

Storage Configuration Advisor validates infrastructure alignment with organizational policies and industry best practices to link configuration problems with associated changes.

- **EMC support matrix validation**

Storage Configuration Advisor automatically downloads the EMC support matrix and check for SAN compliance with EMC E-Lab recommendations.

Figure 25 shows the Storage Configuration Advisor dashboard.

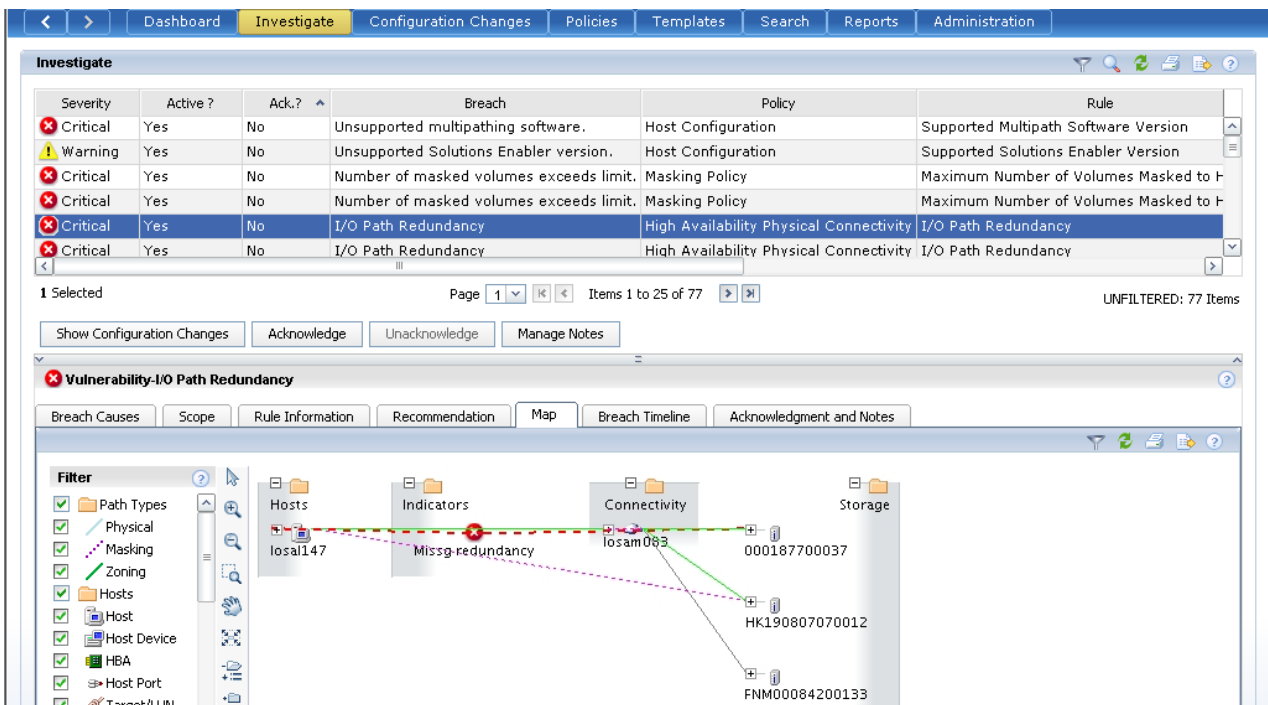


Figure 25. Interface of Storage Configuration Advisor

The Storage Configuration Advisor dashboard provides a high-level snapshot of quantity and impact of change on the environment as shown in Figure 26.



Figure 26. EMC Storage Configuration Advisor dashboard

Appendix C: SQL Server Workload generation tools

Overview

Testing a workload on a new compute, network, or storage platform before migrating the production workload is often desirable. Table 23 outlines some tools commonly used to validate storage subsystems for SQL Server.

Note: Investigate application-specific toolsets before selecting the SQL-only or storage-only toolsets outlined in Table 23. For example, you can use the Microsoft Visual Studio Team Test System (VSTS) to simulate SharePoint workloads end-to-end, including the SQL Server component.

Table 23. Common tools for SQL Server storage subsystem validation

Tool	Advantages	Disadvantages
SQL Profiler	<ul style="list-style-type: none"> Accurately duplicates the SQL Workload Simulates skew Stresses CPU and memory 	<ul style="list-style-type: none"> Relatively complex Requires DBA expertise to execute Requires SQL Admin privilege
IOMeter	<ul style="list-style-type: none"> Simple to set up Tunability Random/sequential Read/write ratio No DBA expertise required Can overlay non-SQL workloads (such as backups) 	<ul style="list-style-type: none"> Difficult to replicate skew Difficult to measure sequentially Does not accurately stress CPU
SQLIO	<ul style="list-style-type: none"> Easy to set up Quickly uncovers infrastructure limits 	<ul style="list-style-type: none"> Cannot approximate production workloads Does not accurately stress CPU
SQLIOSim	Tests I/O stability	Cannot validate disk subsystem performance
Quest BenchMark factory	<ul style="list-style-type: none"> Can simulate TPCC, TPCE, and TPCH-type workloads A commercially supported comprehensive tool that also provides technical support A simulated workload working with SQL Server so that all the system resources will be tested 	<ul style="list-style-type: none"> High license fee Takes time to set up and configure the tool Needs to install SQL server on the server to be tested

Tools introduction SQL Server Profiler

SQL Server Profiler is the most accurate method available to replicate a workload. Database administrators can use SQL Server Profiler to capture traces, and replay those traces on an autonomous system. The primary drawback is that layered workloads occur outside of the context of SQL Server (such as backups).

You can use the Microsoft SQL Server Profiler to capture the Transact-SQL statements sent to SQL Server and the SQL Server result sets for these statements. For information about how to use SQL Server Profiler, see the *SQL Server Profiler* article at the [Microsoft MSDN website](#).

IOMeter

IOMeter is an open source tool used for measuring disk I/O performance. With IOMeter, the administrator can quickly create one or more workers that simulate a custom workload. Typically the workload is measured with a performance monitor or an array-based tool. The drawbacks are the difficulty in measuring (and thus replicate) storage performance characteristics, such as skew and sequentially, and over stressing server-side components, such as CPU and memory. The [IOMeter website](#) has details.

SQLIO

The SQLIO tool was developed by Microsoft to evaluate the I/O capacity of a given configuration. As the name of the tool implies, SQLIO is a valuable tool for measuring the impact of file system I/O on SQL Server performance.

SQLIO is a useful tool for quickly verifying the read and write limits of a disk subsystem. It shares the same drawbacks as IOMeter, but is far less configurable. SQLIO can be run in a manner that creates sequential or random I/O, large or small block I/O, and read or write I/O, but not at the same time. For example, using IOMeter to recreate a database workload that generates 32 KB I/Os, 80 percent random with 75:25 read/write ratio is trivial, using SQLIO to do the same is impossible.

You can download SQLIO at <http://go.microsoft.com/fwlink/?LinkId=115176>.

SQLIOSim

SQLIOSim is a tool designed to verify the stability and not the performance of an I/O subsystem. The tool is useful to verify end-to-end connectivity and stability before deployment or after deployment when disk subsystem or storage network faults are suspected.

For details, and download of SQLIOSim, refer to Microsoft website: <http://support.microsoft.com/kb/231619>.

Quest Benchmark Factory

Quest Benchmark Factory is available to generate databases and workloads for databases, including the SQL Server database. This tool provides TPCC, TPCE, and TPCB workloads with performance recording for running specific workloads. Technical support is available when you purchase the license.

Appendix D: Sample storage designs and reference architectures

Overview

The final step in the SQL Server preproduction deployment phase is to validate the storage for correct configuration and sustain the supported loads.

EMC reference architectures and white papers on the subject of SQL Server storage design are available on the [EMC website](#) and the [Microsoft website](#).

This example demonstrates a detailed design for multi-tier storage of SQL Server OLTP workload in a larger cloud environment on VMAX storage systems and other applications. For details, see the *EMC Virtual Infrastructure for Microsoft Applications White Paper*.

Microsoft SQL Server storage design on VMAX with FAST VP

To maintain flexibility, performance, and granularity of recovery, ensure that the storage sizing and back-end configuration for SQL Server is optimal. This section provides sizing for SQL Server in a FAST VP configuration.

Phase 1—Collect user requirements

Table 24 lists the SQL configuration that meets the user requirements.

Table 24. SQL configuration user requirements

Item	User Equipment
Total Number of users	100,000
Database users per server	20,000; 30,000; 50,000 respectively
Total IOPS	6,000
Number of databases	3
Database profile	Hot/ Warm/ Cold
RPO	Remote < 5 minutes, local = 6 hours
RTO	60 Minutes
Read / Write Ratio	85:15
Backup/Restore required	Yes (Hardware VSS)

Phase 2—Design the storage architecture based on user requirements

EMC recommends that you calculate the disks for the SQL Server to satisfy I/O requirements and then calculate the space requirements. The following is the sizing calculation for this solution.

IOPS calculation

Calculate IOPS as follows:

- Total I/O for 225,00 users is $6000 + 20 \text{ percent} = 6000 + 1200 = 7200$ IOPS
- Calculate the back-end I/O for FAST VP policy requirements for each tier. In this example, the FAST VP sizing is based on the I/O skew of 75 percent SATA, 15 percent FC, and 10 percent flash:
 - Total backend I/O for RAID 1/0 SATA = (10 percent of 7,200) = $(720 \times 0.85) + 2(720 \times 0.15) = 828$
 - Total I/O for RAID 5 FC = (15 percent of 7,200) = $(1080 \times 0.85) + 4(1080 \times 0.15) = 1,566$
 - Total I/O for RAID 5 flash = (75 percent of 7200) = $(5040 \times 0.85) + 4(5040 \times 0.15) = 7,308$
 - The grand total back-end I/O equals 10,224
- SATA disks required to service 808 I/O in a RAID 1/0 configuration is $828/50 \approx 17$ round up to 18 for RAID 1/0
- FC disks required to service 2,088 I/Os in a RAID 5 configuration is $1566 / 130 \approx 12$
- Flash disks required to service 7308 I/Os in a RAID 5 configuration is $7308/1800 \approx 4$

Note: When calculating for performance, the fastest tier needs to service maximum number of I/Os.

- From an I/O sizing perspective, using the previously mentioned policy settings, the following disks would be required for the environment:
 - Eighteen 7.2k, 2 TB SATA drives
 - Twelve 10k, 600 GB FC drives
 - Four 200 GB flash drives

Capacity calculation

- User database size
 - Hot is 200 GB
 - Warm is 300 GB
 - Cold is 600 GB
- Calculate the database LUN size based on the user database sizes:
Database LUN size = <Database Size> + Free space percentage requirement x (20 percent)
 - Hot is $300 + 20 \text{ percent} = 360$ GB
 - Warm is $400 + 20 \text{ percent} = 480$ GB
 - Cold is $700 + 20 \text{ percent} = 840$ GB

- Calculate the Tempdb and log LUN sizes for each of the databases. The log and Tempdb sizes are calculated as 20 percent the size of the database:
 - Log and Tempdb size
 - Hot database is 20 percent of 300 = 60 GB
 - Warm database is 20 percent of 400 = 80 GB
 - Cold database is 20 percent of 700 = 140 GB

The user database log and the Tempdb are on separate LUNs for each database. Based on this, the log LUNs are sized at 120 GB for the hot and warm databases and 140 GB for the cold databases:

- Total database size is the sum of the databases = 2448 GB
- Usable capacity available per 2 TB SATA drive is 1754 GB
- Usable capacity available per 600 GB 10K FC drive is 536 GB
- FAST policy skew used is the total of 75 percent SATA, 15 percent FC, and 10 percent flash
- Capacity for each tier:
 - SATA = $2448 \times 0.75 = 1836$ GB
 - FC = $2448 \times 0.15 = 368$ GB
 - Flash = $2448 \times 0.1 = 245$ GB
- Disk requirement is $\langle \text{Total capacity} \rangle / \langle \text{Usable Capacity} \rangle$
- Disks required for each tier:
 - SATA (mirrored) = 4
 - FC (RAID5 3+1) = 4
 - Flash (RAID5 3+1) = 4

Note: When calculating for capacity, the slowest tier needs to host the majority of the data.

- From a capacity sizing perspective, using the policy settings mentioned above, the following disks are required for the environment:
 - Four 7.2k 2 TB SATA drives
 - Four 10k 600 GB FC drives
 - Four 200 GB flash drives

The best configuration is based on both I/O and capacity requirements, as shown in Table 25.

Table 25. Best configuration based on both I/O and capacity requirements

	1 TB (200GB, 300GB, 500GB) SQL Server database
Number of disks required to satisfy both I/O and capacity	18×7.2K 2 TB SATA drives 12×10K 600 GB FC drives 4×200 GB flash drives
Thin LUN sizes (Database)	Hot is 360 GB Warm is 480 GB Cold is 840 GB
Thin LUN sizes (Log)	Hot is 120 GB Warm is 120 GB Cold is 140 GB

Building-block design approach for data warehouse

This solution describes a building-block design storage of a data warehouse for flexibility and performance scalability. For details, refer to the *SQL Server 2012 Data Warehouse White Paper*.

Building-block design considerations

The infrastructure supporting the data warehouse, including the server, network, storage, and application, must provide a robust, powerful, and flexible solution.

This design is for data warehouses in a virtualized environment that provides predictable performance. You must consider the following criteria for this design.

Proportional and predictive bandwidth

- **Design for bandwidth, not for database capacity.** In this example, to achieve the completion of all DSS queries in the test suite in a 12- to 14-hour window, the desired bandwidth is 100 MB/s for a 500 GB database, 200MB/s for a 1 TB database, or 400 MB/s for a 2 TB database.
- **Check disk performance.** For a DSS workload with sequential 64 K read-only I/O, the 10 K, 600 GB SAS disks used in this example provide average IOPS of 320 and bandwidth of 20 MB/disk. Bandwidth can be calculated for a given I/O size with the IOPS:
Bandwidth = Average I/O size × IOPS
- **Calculate the building block storage requirements.** With a DSS workload of read only I/O, on a RAID 5 (4+1) configuration, the 10K 600 GB SAS disks needed would be:
Disk number = required bandwidth / bandwidth per disk
 - For the 500 GB building-block with bandwidth of 100 MB/s, five 10 k 600 GB SAS disks would be required.

- For a 1 TB database building-block with 200 MB/s bandwidth, 10 disks would be required.
- For a 2 TB building-block with 400 MB/s bandwidth, 20 disks would be required.

Virtual machine scalability

The virtual machine scalability requirements are:

- Virtual machine resources including vCPU and memory allocation should be part of the building block.
- The building block should be able to scale up (adding a block into the same virtual machine) and scale out (adding the building-block to another virtual machine) all with minimum performance degradation.

Sufficient resource

Confirm the building-block design of the disks and memory are sufficiently used as follows:

- **Disk utilization** sufficiently uses the disk resource with room for any possible peak disk activities.
- **System memory** utilization supports the designed workload with anticipated peak load activities.
- **vCPU processor** utilization supports the designed workload and any anticipated peak load activities.
- **Tempdb design** has sufficient capacity and performance to support the query workload of the database. The DSS workload has relative high demand for Tempdb.

Balanced disk utilization

Build database LUNs across as many buses as possible to avoid an unbalanced workload. This allocation can have potential high-availability benefits.

Building-block design details

The building blocks listed here are designed as follows:

- Targeted bandwidth is 100 MB/s per LUN (R5 4+1 10K 600 GB SAS disks). Use more disks for the building block if you want higher bandwidth.
- In this example the database size scaled for each LUN (R5 4+1) is 500 GB. For a 1 TB database, two LUNs (10 disks) are created for the database files. For a 2 TB database, four LUNs (20 disks) are created.
- To support the targeted bandwidth of 100 MB/s per LUN, assign a minimum of two vCPUs and 8 GB memory in proportion with the building block. Or one vCPU and 4 GB memory per 50 MB/s of bandwidth.

Figure 27 shows the three building blocks used in this example based on this design principle. Table 26 lists the details for the three building blocks.

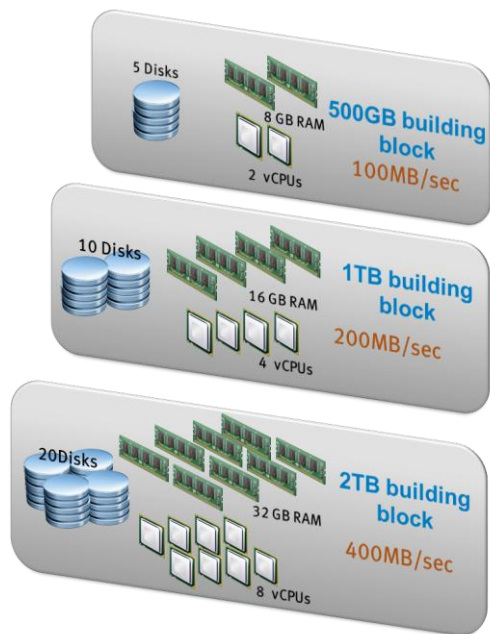


Figure 27. Three tested building blocks

Table 26. The building-block configurations

Configuration	500 GB building block	1 TB building block	2 TB building block
Database size	500 GB	1 TB	2 TB
Targeted bandwidth (MB/s)	100	200	400
Database LUN design	1 x 2 TB data LUN 8 x 80 GB data file	2 x 2 TB data LUN 8 x 126 GB data file	4x 2 TB data LUN 16 x 126 GB data file
Log design	1 log LUN (5 GB log file)	1 log LUN (12 GB log file size)	1 log LUN (12 GB log file size)
Tempdb design	1 data LUN(1x 100 GB data file) 1 log LUN (2 GB log file)	1 data LUN (2x 100 GB data file) 1 log LUN (2 GB log file)	1 data LUN (4x 100 GB data file) 1 log LUN (2 GB log file)
Disk configuration	5 SAS disks	10 SAS disks	20 SAS disks
Memory (GB)	8	16	32
vCPU (2.4 GHz)	2	4	8

Note: Log LUNs in a data warehouse environment are not heavily utilized, so multiple building blocks on the same virtual machine can potentially share the same log LUN.

Table 27 defines the minimum requirements for the building blocks and virtual machines in this example.

Table 27. Building-block design

Per virtual machine	Memory (GB)	CPU (core)	Number of RAID 5(4+1) disks
Minimal	16	4	5
/TB	16	4	10
/(100 MB/s)	8	2	5
500 GB (100 MB/s)	8	2	5
1 TB (200 MB/s)	16	4	10
2 TB (400 MB/s)	32	8	20
4 TB (800 MB/s)	64	16	40
6 TB (800 MB/s)	96	24	60

Deploying building blocks

Figure 28 shows the following two ways to deploy building blocks:

- Scale-up design that puts the building blocks on the same virtual machine.
- Scale-out design that puts each building block into a different virtual machine.

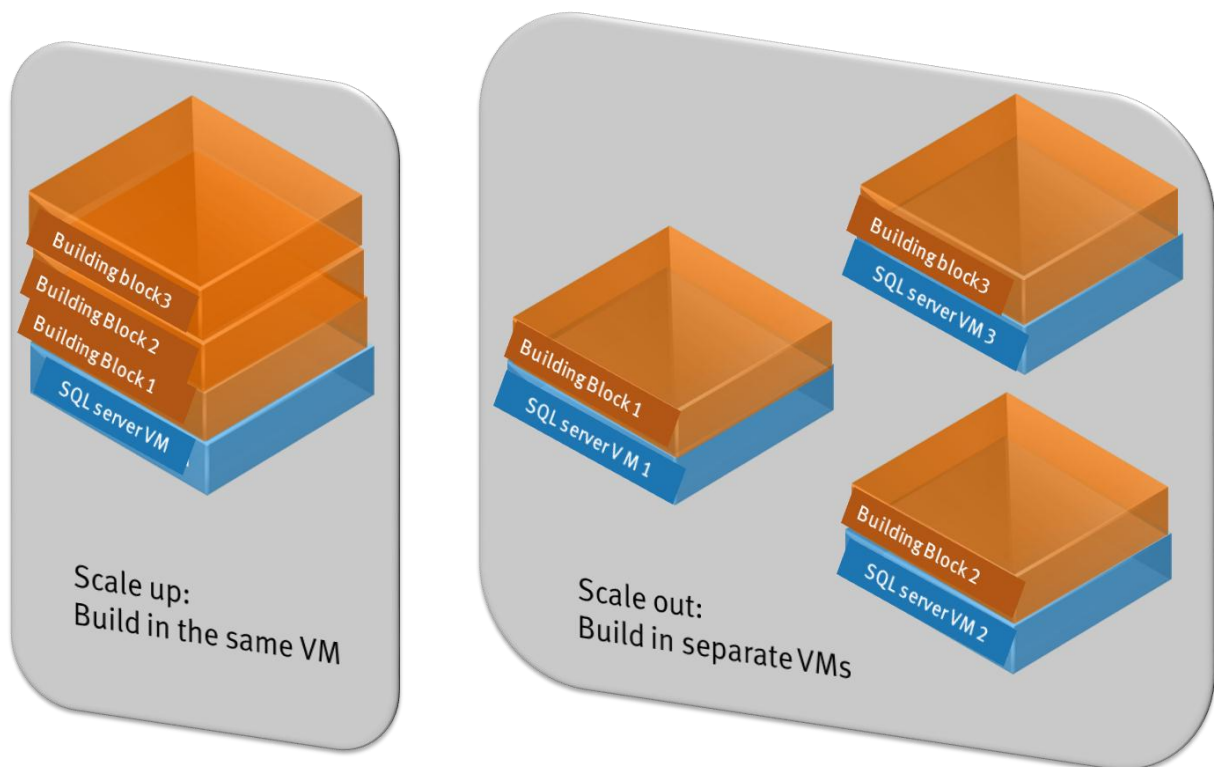


Figure 28. Scale-up and scale-out building blocks

Scale-up design

The scale-up design can potentially save the license cost for the OS.

Because the vCPUs and memory resources are built-in to the building-block design, these resources grow proportionally with the building-block deployment for the same virtual machines required to support that building block.

Scale-out design

In a scale-out deployment of building blocks, the vCPU and memory capacity of the ESXi server must be considered to support the number of building blocks desired.

The system resources such as vCPU and memory are within the building block. Adding a building block in a scale-up (to the same virtual machine) or a scale-out (to a separate virtual machine) requires the same resources unless built below the minimum requirement for a virtual machine. Thus, when having small building blocks (such as a database less than 1 TB with 200 MB/s bandwidth), it might be better to use the scale-up model to reduce waste of virtual machine system resources such as memory and CPU.

In a midsized environment such as this solution, organizations should carefully consider which approach best suits them.

SQL Server virtual machine and LUN allocation design

Table 28 and Figure 29 show the virtual machine configuration and disk assignment for different databases used in this example. We tried various ways to design and deploy a building block with reasonable performance.

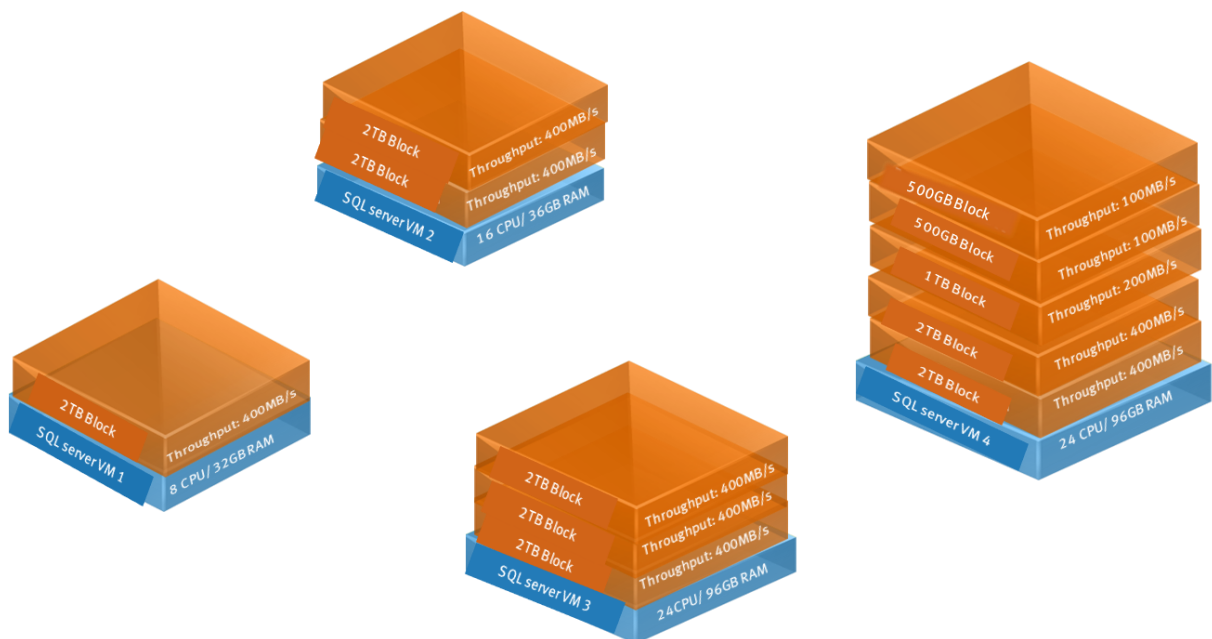


Figure 29. Solution building block deployment

Note: In one of the virtual machines, five databases with various size and bandwidth requirements were all running heavy workload at the same time. Thus, they would compete for resources and might have contentions at peak time. Even in this case, the SQL Server in the virtual machine maintains good performance.

Table 28. Solution building-block deployment on the virtual machine and ESXi server

Host	Allocation	CPU (core)	Memory (GB)	Database size (TB)	Number of disks	Bandwidth (MB/S)
ESXi 01 (40 cores, 256GB RAM)	VM1	16	64	2	20	400
	VM2	8	32	2	20	400
ESXi 02 (80 cores, 512GB RAM)	VM3	24	96	2	20	400
				2	20	400
				1	10	200
				0.5	5	100
				0.5	5	100
	VM4	24	96	2	20	400
Total	4 virtual machines	72	288	18	180	3600

SQL Server protection solution

This solution implemented continuous protection for SQL Server by using EMC RecoverPoint with Replication Manager and vCenter SRM. For details, refer to the *Continuous Data Protection for Microsoft SQL Server Enabled by EMC RecoverPoint, EMC Replication Manager, and VMware White Paper*.

EMC RecoverPoint

This section describes the configuration required for EMC RecoverPoint when implementing protection.

Local replication process (CDP)

Figure 30 shows the RecoverPoint Continuous Data Protection (CDP) process that synchronously replicates data from the production (source) volumes to the local target volumes, while maintaining reversible recovery through journaling storage.

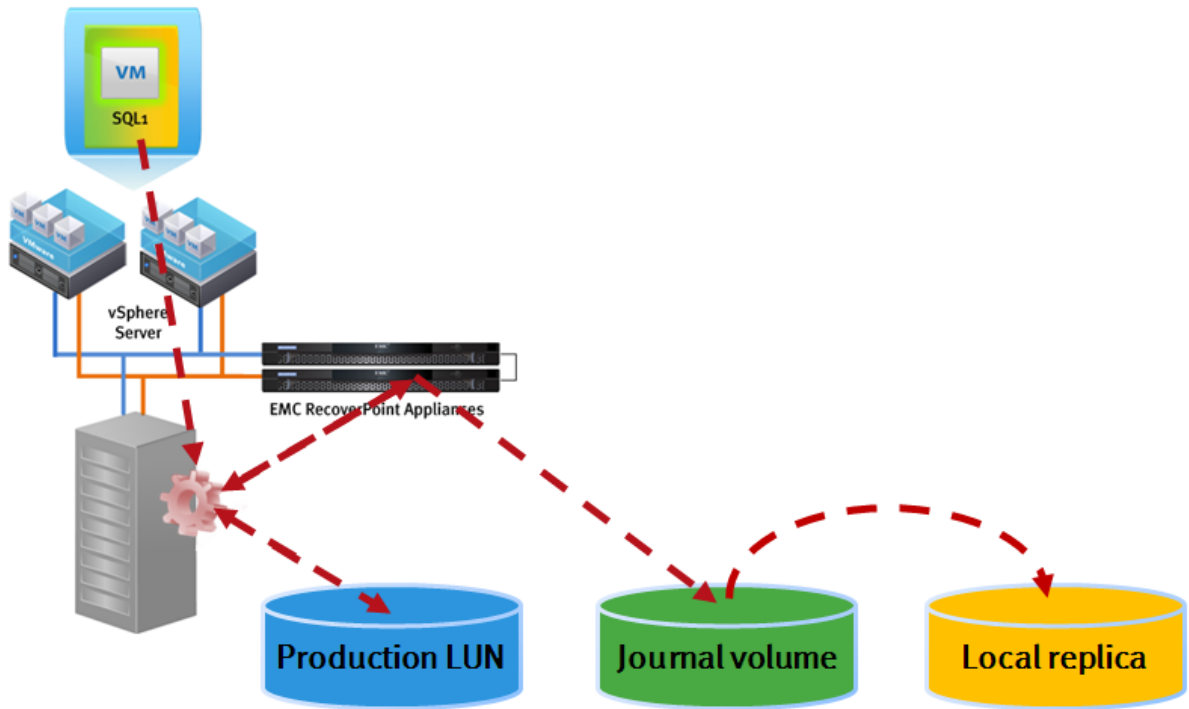


Figure 30. EMC RecoverPoint local replication (CDP) data flow

Remote replication process (CRR)

Table 22 shows the RecoverPoint Continuous Remote Replication (CRR) process that replicates blocks of data to remote site storage array. The data is replicated either synchronously over a Fibre Channel connection up to 200 kms/4 ms or asynchronously over an IP connection, which in this example tested up to 64 ms/6,400 km round trip.

Figure 31 shows a comparison for CDP and CRR process with remote RecoverPoint replication.

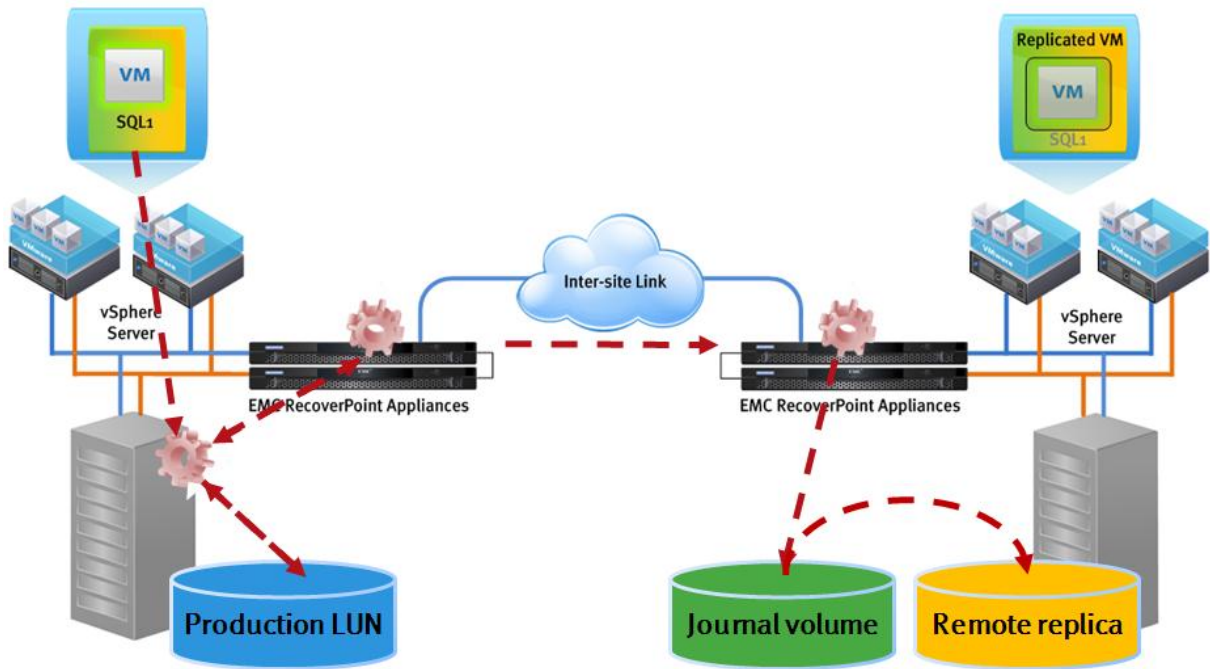


Figure 31. RecoverPoint remote replication (CRR) data flow

Table 29. The CDP and CRR process for RecoverPoint local and remote replication:

Operations	CDP (EMC RecoverPoint local replication) data flow	CRR (RecoverPoint Remote replication) data flow
A write to LUN protected by RecoverPoint	Write intercepted by the RecoverPoint splitter.	Write intercepted by the RecoverPoint splitter.
Splitter splits the write	Simultaneously sends it to the production volume and to the local RPA	Simultaneously sends it to the production volume and to the local RPA.
Acknowledge	Writes are acknowledged back from the RPA and production LUN immediately.	Asynchronous replication: Writes are acknowledged back from the RPA and production LUN immediately. Synchronous remote replication: Acknowledgment (ACK) is made when the write has been received at the remote site.
Time stamp and bookmark writes to journal	The RPA writes data to the journal volume along with time stamp and bookmark metadata.	Local RPA bundles the write with other writes, sequences and time stamps the write. The package is compressed and transmitted with a checksum for delivery over IP to the remote RPA.

Operations	CDP (EMC RecoverPoint local replication) data flow	CRR (RecoverPoint Remote replication) data flow
	N/A	Remote RPA receives the package, verifies the check sum to ensure no corruption in transmission, then decompresses the data.
	N/A	Remote RPA writes the data to the journal volume.
Complete	After the data is safely stored in the journal, write-order-consistent data is distributed to the local replica.	After the data is written to the journal volume, it is distributed to the remote volumes. Write order is preserved during this distribution.

Consistency groups

A consistency group is a logical container within RecoverPoint, which ensures all devices within that consistency group are consistent (write-order fidelity) with each other. RecoverPoint version 3.3 supports up to 128 consistency groups.

For SQL Server Virtual machines, the *Resource Allocation Priority* policies are defined based on volume requirements:

- **OS/Page File** –Operating system and page file volumes with natural relationship are kept together. With a very low change rate, the Resource Allocation Priority is set to **Normal**.
- **Tempdb/Systemdbs** –Tempdb and Systemdb are replicated to bring up all virtual machines at the remote site. So because the Systemdb does not change much and Tempdb is recreated whenever SQL Server instance is restarted, this is expendable so set the Resource Allocation Priority to **Low**.
- **Data/Logs**–User data is the most important in the environment, so set the Resource Allocation Priority to **Critical**.

Figure 32 shows the relationship between the Windows NTFS volumes, their relevant consistency groups, and what consistency group policy settings were defined for both local and remote policies.

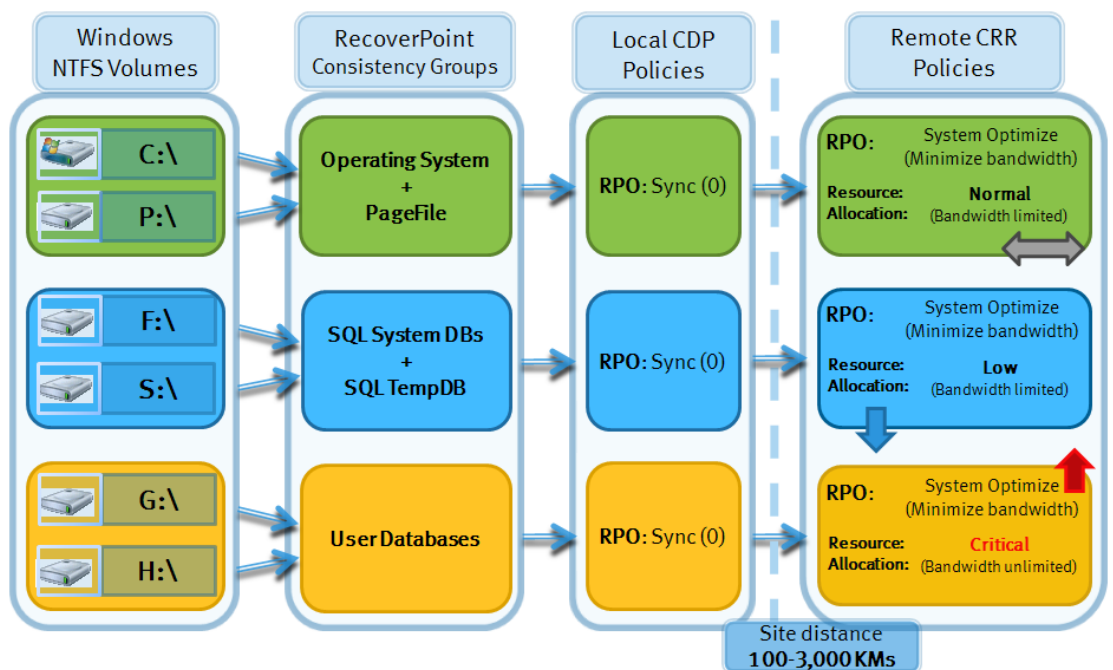


Figure 32. Windows volumes to RecoverPoint consistency group mapping

When using CLR for each consistency group, three journals were set up: two at the production site to support both CDP and CRR and one at the recovery site for CRR.

Groups sets

The RecoverPoint Group Sets feature allows consistent bookmarking across multiple consistency groups. A Group Set can be made to contain consistency groups for a single virtual machine.

In order to access a copy of the replicated data, enable Image Access on each of the consistency groups for the time required (Figure 33). All three consistency groups are rolled back to the same target image to ensure volumes at the disaster recover site are fully consistent TPCE1 virtual machines and the server can be restarted in a crash-consistent state.

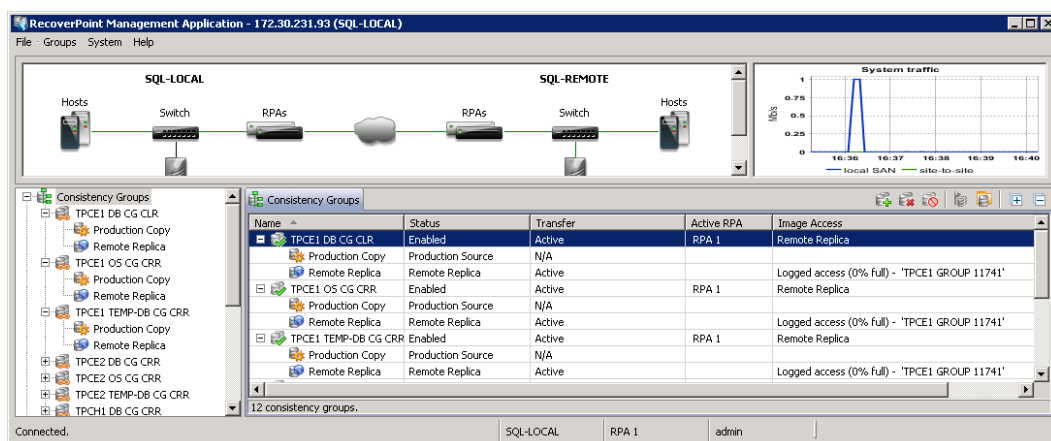


Figure 33. RecoverPoint Management Console – consistency groups

Journal sizing – protection windows

A critical consideration is the sizing of RecoverPoint journals. They must have both performance and capacity characteristics in order to handle the total write performance and store all the writes of the LUN being protected. The two most important questions are:

- What change rate does the source LUN generate?
- What retention window is required?

To calculate the journal capacity, the rate of change must be measured on the production LUNs. Perfmon counters are set on each of the SQL Servers to capture the write bandwidth in megabytes per second (MB/s). Data per second can be found with the Unisphere Analyzer, while the **Stats** tab gives a point-in-time window to see what each storage processor is doing.

The Journal Volume Sizing formula is:

$$\text{Journal size} = \frac{(\text{data per second}) \times (\text{required rollback time in seconds})}{(1 - \text{target side log size}) \times 1.05}$$

Twenty percent of the journal must be reserved for the target side log and five percent for internal system needs.

For example, to support a 24-hour rollback requirement (86,400 seconds), with 5 megabits per second (Mb/s) of new data writes to the replication volumes in a consistency group, the calculation is:

$$\frac{5 \times 86,400}{(1 - 0.2) \times 1.05} = 567,000 \text{ Mb} = 69.213 \text{ GB} (\sim 70 \text{ GB})$$

For the solution, all journals were sized to enable RecoverPoint to roll back at least seven days.

Integrating RecoverPoint with VMware vCenter

The vCenter Servers view displays data from the vCenter Server through the RecoverPoint graphical user interface (GUI). In addition to displaying ESX Servers and all their virtual machines, datastores, and RDM drives, the vCenter Servers view also displays the replication status of each volume. The protection status of every virtual machine is measured multiple times per hour. This window is updated when a new virtual machine is created or the protection status of a virtual machine changes. The vCenter Servers view is for monitoring only (read-only).

For example, as shown in Figure 34, the RecoverPoint Management Application shows that all relevant volumes for TPCE1 and TPCH2 are successfully replicated. The respective consistency groups, the copy being replicated, and the associated replication sets are shown in Figure 34.

Name	IP	Consistency group	Copy	Replication set	Datastore
TCE-SQL-TPCE1-75	172.30.231.74				
naa.600601604538270092dc291e48f2df11		TPCE1 TEMP-DB CG CRR	Production Copy	Temp DB	F600'S VM TPCE1-75K TEMP DB & LOGS
naa.600601604538270092dc291e48f2df11		TPCE1 OS CG CRR	Production Copy	OS	F600'S VM OS TCE-SQL-TPCE1-75
naa.600601604538270092dc291e48f2df11		TPCE1 DB CG CLR	Production Copy	LOGS	F600'S VM SQL TPCE1 LOGS
naa.6006016045382700baba839d2e5df11		TPCE1 DB CG CLR	Production Copy	DB	F600 VM SQL OLTP 75K
naa.6006016045382700c4904b14250de011		TPCE1 OS CG CRR	Production Copy	Page File	F600'S VM PAGE FILE TCE-SQL-TPCE1
naa.6006016045382700e02100e24cf2df11		TPCE1 TEMP-DB CG CRR	Production Copy	SQL SYSTEM	F600'S VM SQL SYSTEM TPCE1
TCE-SQL-TPCH2	172.30.231.77				
naa.60060160453827005418e28534f2df11		TPCH2 DB CG CRR	Production Copy	DB2	F600 VM SQL TPCH2 DB2
naa.600601604538270094011d5434f2df11		TPCH2 DB CG CRR	Production Copy	DB1	F600 VM SQL TPCH2 DB1
naa.6006016045382700960991d24bf2df11		TPCH2 TEMP-DB CG CRR	Production Copy	SQL System	F600'S VM SQL SYSTEM TPCH2
naa.6006016045382700ba8ae4fe3ff2df11		TPCH2 TEMP-DB CG CRR	Production Copy	Temp DB	F600'S VM TPCH2 TEMP DB & LOGS
naa.6006016045382700e69a6372b7f1df11		TPCH2 OS CG CRR	Production Copy	OS	F600'S VM OS TCE-SQL-TPCH2
naa.6006016045382700f6db8e714af2df11		TPCH2 DB CG CRR	Production Copy	Logs	F600'S VM SQL TPCH2 LOGS
naa.600601607d1925006e326e36371fe011		TPCH2 OS CG CRR	Production Copy	Page File	F600'S VM PAGE FILE TCE-SQL-TPCH2

Figure 34. vCenter Servers view in RecoverPoint Management Application

RecoverPoint failover

When a virtual machines crashes, RecoverPoint failover process can be used to failover to bring up the disaster recovery site.

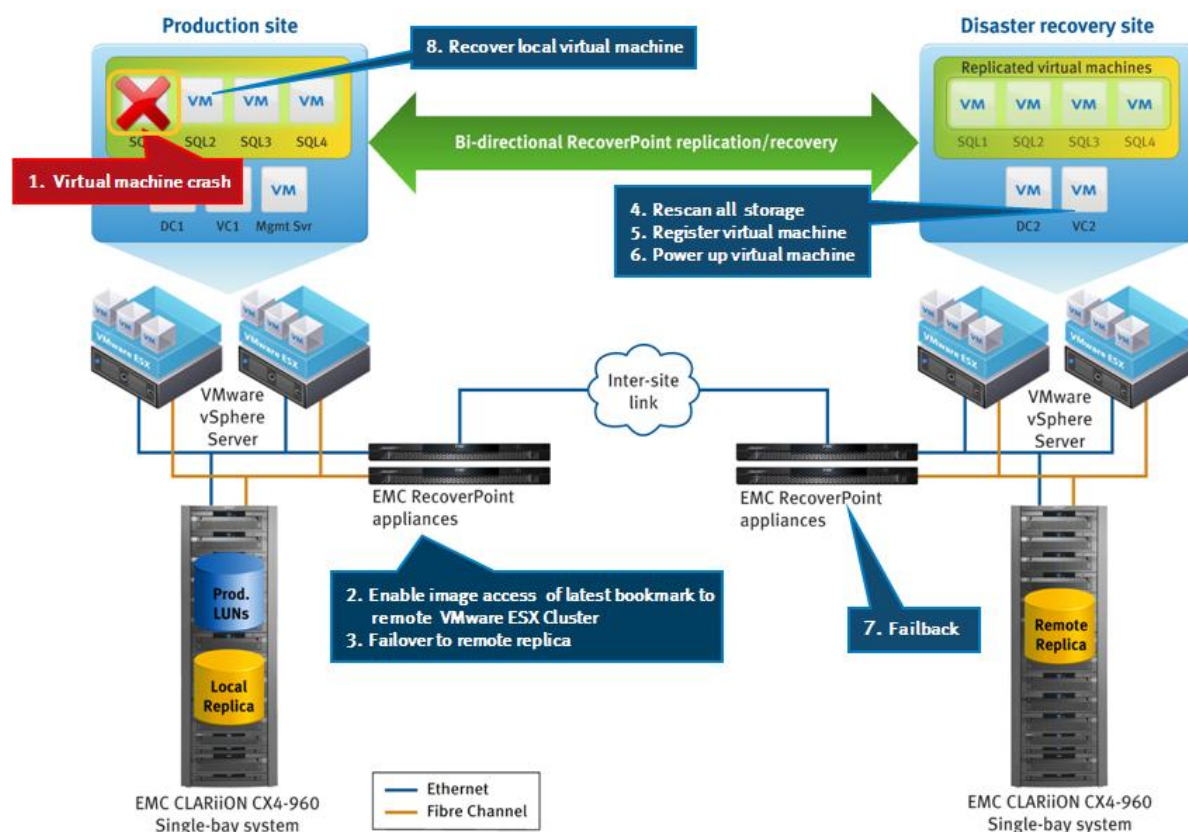


Figure 35. RecoverPoint failover process

In this scenario (as shown in Figure 35), use the following steps in Table 30 for failover when a virtual machine crashes at the production site.

Table 30. Failover steps

RecoverPoint failover process	Details
Enable image access to the latest bookmark.	This provides read/write image access of the CRR copy to the remote ESX Servers and allows the mount of the VMFS volumes on the remote vCenter.
Confirm the failover is set to the remote replica.	In the RecoverPoint Management Application or with the Unisphere Management Console, choose failover to the remote replica .
Confirm the remote storage is fully accessible to ESX Server.	Rescan all storage on the remote ESX Server through the remote vCenter console.
Register the remote site virtual machine.	Right-click on the VMX file in the OS LUN VMFS datastore to register the virtual machine and choose to Inventory the virtual machine .
Start up the virtual machine.	Connect the vNIC to the network to re-enable IP access to the virtual machine and database.

You can perform failback by shutting down the virtual machine on the remote site and repeating the above steps for the production site.

This solution allows for full portability of your SQL Server instances between the sites.

Different subnet

If the virtual machine is being failed over to a vSphere cluster on a different subnet (for example, from 10.10.10.x to 10.20.20.x), a distributed switch needs to be created on the production ESX cluster with the same properties as the actual distributed virtual switch for the disaster recovery site. In order to configure for failover to a different subnet, you would assign a vNIC on the dummy switch to the SQL Server virtual machines in production.

As shown in Figure 36, a dummy switch is created at the production site vCenter Server. The virtual machine is then configured with a second vNIC on the dummy virtual disaster recovery switch. This allows the virtual machine to fail over seamlessly to the remote site without any additional network configuration for the disaster recovery site.

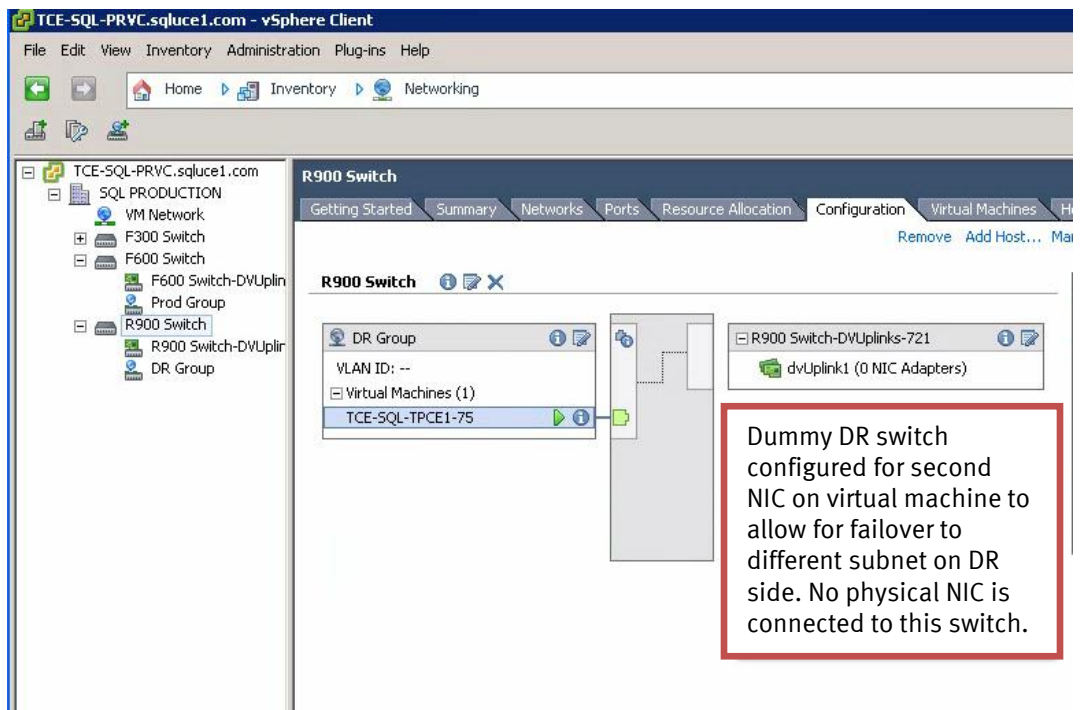


Figure 36. vNIC is configured and connected

When the virtual machine is failed over to the disaster recovery site, this NIC will be connected to the specified (sqluce1.com) network, as shown in Figure 37. The production site vNIC will then have an unidentified network because the production switch configured at the disaster recovery site is only there for configuration purposes and is not live on the network.

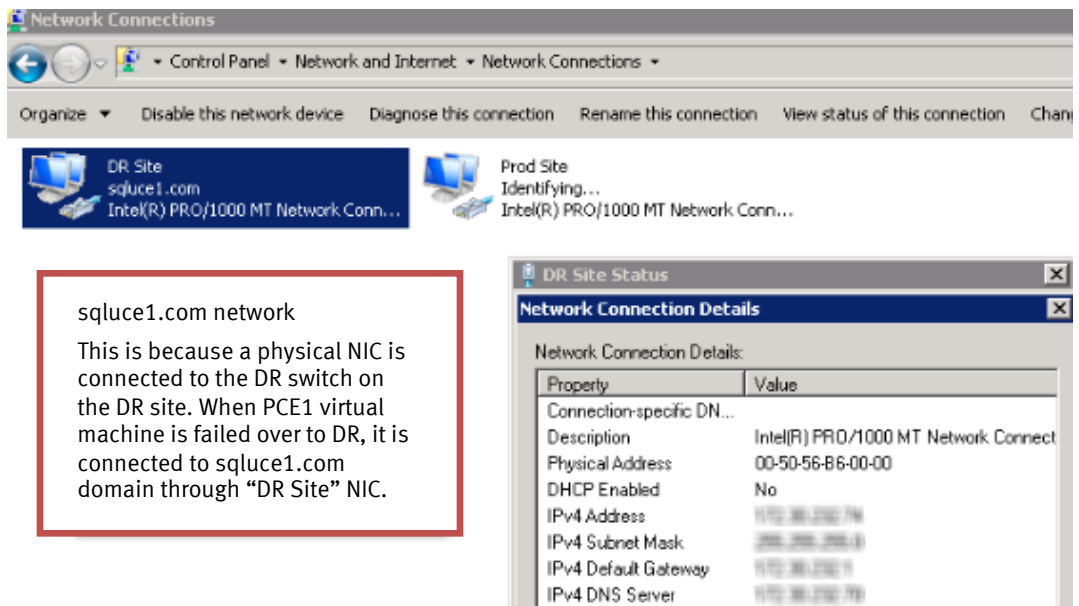


Figure 37. Disaster recovery site connected to sqluce1.com network

EMC Replication Manager

This section describes the configuration required for Replication Manager when implementing protection.

Integrating Replication Manager with SQL Server

To support SQL Server, Replication Manager uses the SQL Server Virtual Device Interface (VDI) snapshot API for online, rapid application-consistent snapshots of very active enterprise-class SQL Server databases with negligible host overhead.

Replication Manager enables you to perform the following tasks with a simple wizard-driven interface:

- Specify which instances, databases, and corresponding filegroups to replicate.
- Ensure that the data can be replicated safely and quickly.
- Return the database to normal operation after creating the replica.
- Mount or recover a database on another host for other operations, such as testing, reporting, or data mining.
- Quickly recover a database on the production host if data corruption occurs.

Application sets and jobs for Microsoft SQL Server

Replication Manager uses the concept of application sets as containers to define what data to protect (for example, Database 1) and jobs as a way to protect that data (for example, RecoverPoint bookmark image).

Microsoft SQL Server application consistency

In this example, the Replication Manager jobs were configured to protect SQL Server user databases by setting the **Replica Type** option to **Full, Online with advanced recovery using VDI**. This option replicates the entire database and transaction log. This replica type is typically used when the replica is considered a backup of the

database or when the replica is mounted for a third-party product to create a backup of the database.

To bring the database forward to a point in time that is newer than the replica, this replica type allows you to restore transaction logs, assuming you have backed up those transaction logs.

Replication Manager uses VDI-enabled snapshots to create this replica type, guaranteeing application-consistent data.

Note: The system databases (master, MSDB, and model) must not be located on the same volume as user databases. Microsoft SQL Server does not support VDI and snapshot technology to restore system databases.

Configuring Replication Manager to communicate with vCenter and RecoverPoint

Replication Manager can replicate, mount, and restore a VMFS datastore at the LUN level. The Replication Manager software or agent does not need to be installed on the ESXi server or virtual machine. All operations are performed through the vCenter and Replication Manager VMware proxy host that can be either a physical or virtual host. The proxy host must be registered with the Replication Manager server with the credentials for vCenter management. The Replication Manager VMware proxy host communicates with vCenter over port 443. Replication Manager can map the associated VMFS volumes to the LUNs.

The proxy host shares the same virtual machine as the Replication Manager Server in this example. Replication Manager can also discover the LUNs replicated by RecoverPoint. Communication to RecoverPoint is done through the Replication Manager agent installed on the production and mounted virtual machines.

Logical Volume Manager resignaturing

VMFS replication requires that Logical Volume Manager (LVM) resignature be enabled on both the production and mount ESX Servers. LVM resignature allows VMware to write a new signature to the LUNs when necessary. This switch must be enabled for Replication Manager so that VMFS can be made visible on the replicated LUNs to the ESX Server. This setting should also be enabled on the production ESX Server in order to restore to that ESX Server at any time.

The following command must be issued on ESX Servers used to mount replicas:

```
esxcfg-advcfg -s 1 /LVM/EnableResignature
```

For more information on this topic, refer to the *EMC Replication Manager Administrators Guide*, *VMWare Setup* section.

Discovering the RecoverPoint appliance and storage array

This solution uses the storage RecoverPoint splitter, which enables the storage array to discover the RecoverPoint storage within Replication Manager. You need to enter your storage credentials to complete this discovery task.

After you configure the credentials for at least one Replication Manager host agent, a storage array discovery operation also detects the RecoverPoint splitter.

Recovering a SQL Server user database

In this example, EMC simulated a disaster of a live production database and tested the solution by recovering to a specific point in time.

For the test, a table from an OLTP database is deleted at 14:16:00. This is a critical table to the functionality of the database called the Accounts Permissions Table. Users cannot access data without it, so at 14:16:00, the business in which the database was serving goes down. Then, the entire database was deleted to simulate human error and the entire database is lost at 14:16:15.

Replication Manager is the only interface required to recover the database, because it coordinates all operations across all levels of the solution stack, including SQL Server, Windows Server, VMware, EMC storage, and RecoverPoint, to orchestrate the recovery process. The wizard-driven interface is useful when recovering a business-critical transactional database, ensuring all best practices are followed for a successful restore. This solution effectively restored the user database with the CRR copy from the remote site, and wound the clock back to just one second before the disaster occurred, at 14:15:59.

Replication Manager accessed an image of the database at that exact time specified and recovered the database, which resulted in an RPO of one second. EMC chose to recover all files and filegroups for the database.

Replication Manager took 3 minutes and 26 seconds to finish the recovery process, leaving the database detached, which allows the DBA to attach the database with the ensured integrity before allowing users access.

Once the database was attached, online users could access data at 14:21:45, and the business unit was backed up.

This solution gave an RPO of one second and an RTO of less than four minutes. The level of recovery is extremely powerful, allowing you to commit to strict service level agreements (SLAs). This enables easy and quick recover business-critical, highly transactional OLTP databases with minimal steps.

VMware vCenter SRM

This section describes the configuration for VMware vCenter SRM when implementing protection.

Integrating vCenter SRM with RecoverPoint

vCenter SRM reduces the RTO for disaster recovery and relies on block-based replication to reduce the RPO for disaster recovery. The RecoverPoint SRA is used to map the vCenter SRM requests into the appropriate RecoverPoint actions.

vCenter SRM and RecoverPoint automate the virtual machines recovery process, which makes it as simple as pressing a single button. The user has no interaction with the RecoverPoint console; instead, vCenter SRM automates the whole failover process. The integration between RecoverPoint and vCenter SRM is controlled by the RecoverPoint Storage Replication Adapter (SRA).

RecoverPoint is responsible for replicating all changes from the production LUNs to the remote replicas at the disaster recovery site. The RecoverPoint SRA is installed on the same servers that are running the vCenter Server and the vCenter SRM plug-in at

the production and disaster recovery sites. RecoverPoint SRA supports vCenter SRM functions, such as failover and failover testing, by using RecoverPoint for replication.

Configuring the consistency group for management by vCenter SRM

After the consistency group is created and vCenter SRM is installed, you need to configure the consistency group to be managed by vCenter SRM. Use the policy settings in the RecoverPoint Management Application to do this, as shown in Figure 38.

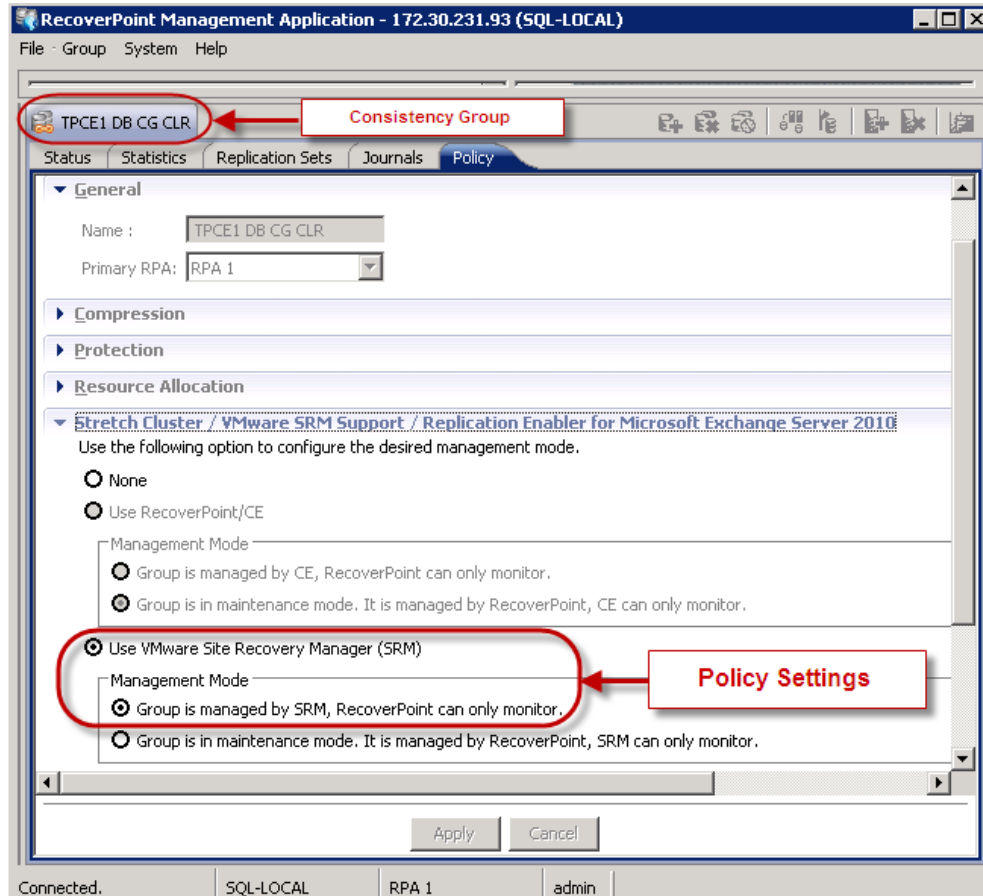


Figure 38. Configuring the consistency group for management by vCenter SRM

vCenter SRM solution protection

In this example, vCenter SRM is protecting the SQL Server virtual machines. Replication Manager is required on the production site for local protection, mount, and restore. The disaster recovery site contains its own vCenter and Active Directory virtual servers, so there is no requirement to replicate these.

Figure 39 shows how vCenter SRM protects the SQL Servers using RecoverPoint integration by automating the required steps.

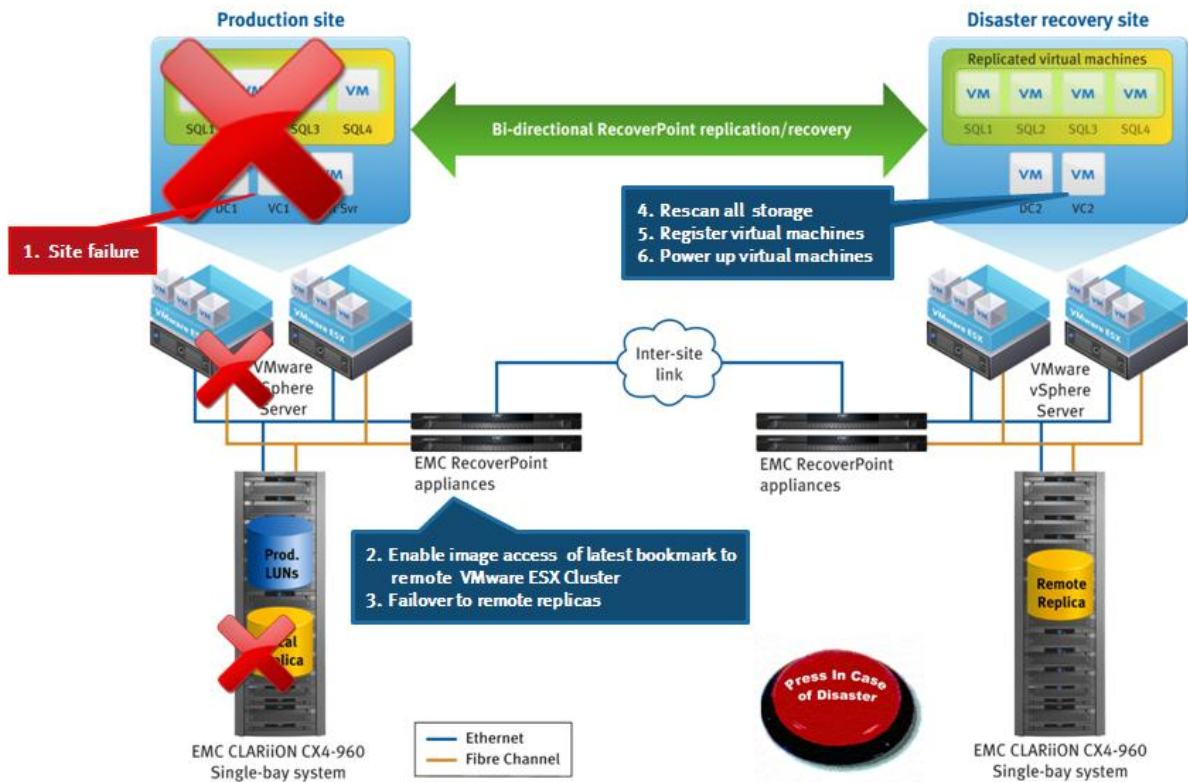


Figure 39. vCenter SRM protection procedure for production site

vCenter SRM requires configuration on both the production and recovery sites.

The production site requires the following configuration:

- Connection to establish vCenter SRM communication between vCenter Servers
- Array managers to detect replicated devices
- Inventory mappings for site-specific folder, network, and resource mappings
- Protection groups to organize virtual machines on their respective datastores for recovery

The recovery site requires that you configure a recovery plan by creating an automated runbook of the recovery process.

Configuring vCenter SRM protection groups

A RecoverPoint consistency group is a data set of SAN-attached storage volumes at the production site and disaster recovery site. A vCenter SRM protection group is a group of virtual machines that fail over together (during testing and actual failover).

When vCenter SRM performs a failover, it instructs RecoverPoint to operate on the LUNs of all virtual machines in the protection group. However, RecoverPoint uses consistency groups to define groups of LUNs that are replicated together.

After successfully configuring the connection, array managers, and inventory mappings, you need to configure the protection groups, as shown in Figure 40.

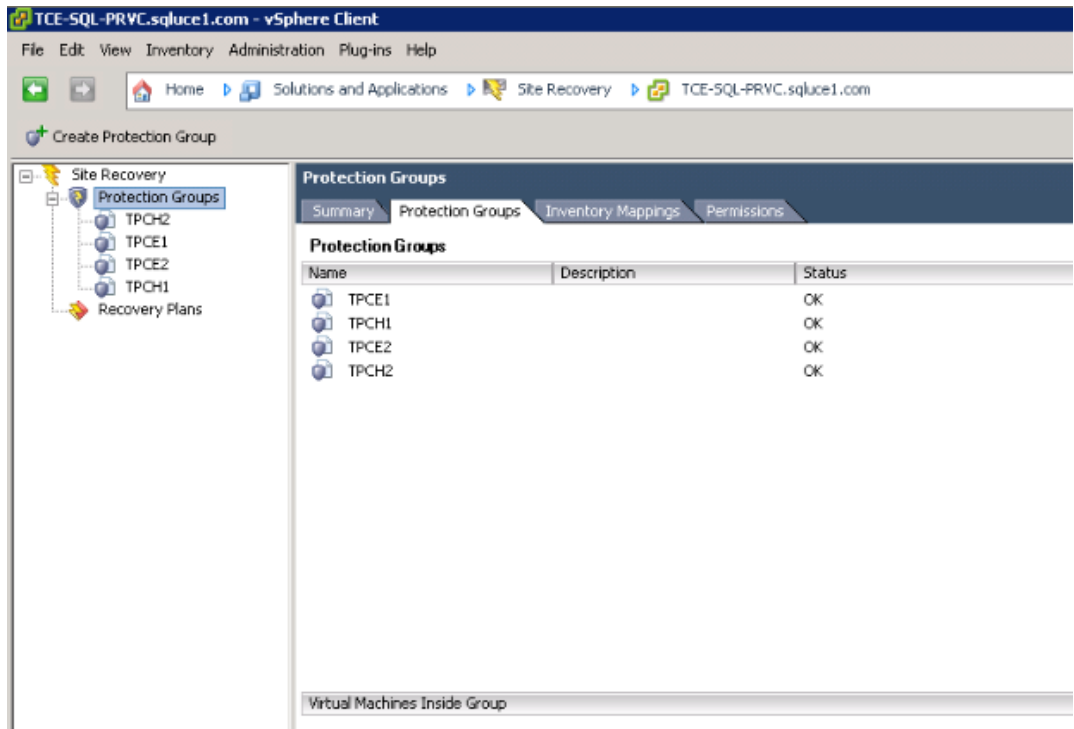


Figure 40. Configuration of protection groups in vSphere Client

For this solution, four individual protection groups were created, two for each of the TPCE virtual machines and two for the TPCH virtual machines. Within each protection group, you can specify the recovery priority for each virtual machine.

Modifying the startup priority of a virtual machine

You might not want all virtual machines to restart simultaneously on recovery, therefore, EMC configured the TPCE virtual machines for high priority, as they are the most critical to the business (as shown in Figure 41), and left the TPCH virtual machines on a low priority setting.

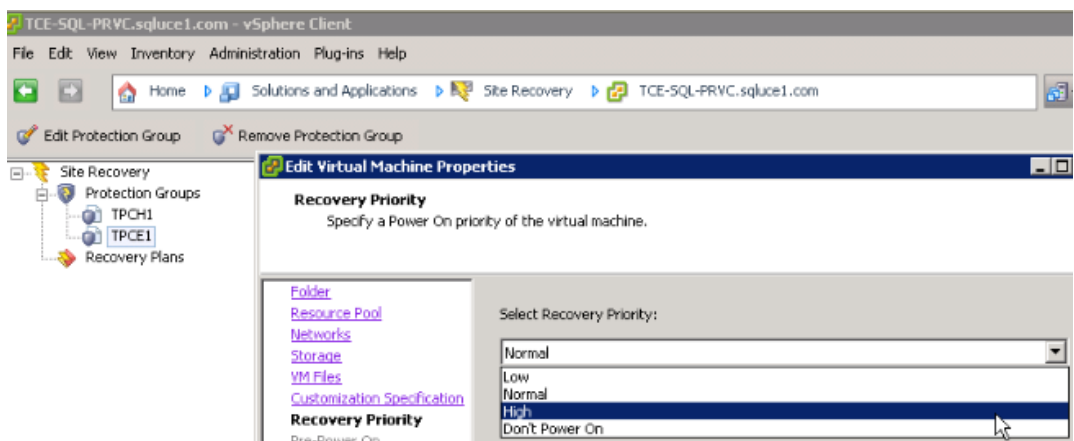


Figure 41. Selecting recovery priority for virtual machines

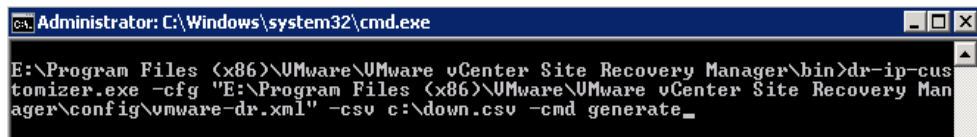
Customizing recovery site IP addresses

When failing over to a different data center, some adjustments are required to the host IP settings for infrastructure differences. When failing over an entire configuration, this can involve updating the settings for multiple virtual machines.

vCenter SRM provides a bulk IP customization utility (`dr-ip-customizer.exe`) for automatically updating IP settings for recovered virtual machines. The utility generates a CSV file containing the IP settings for all virtual machines that are configured for vCenter SRM failover. You can edit this file to specify the recovery site IP settings and then run the utility again to upload the new settings to the recovery site vCenter server.

For this solution, the utility was used to update the recovery site IP settings as follows:

1. Log on to the vCenter server at the recovery site.
2. Run the `dr-ip-customizer.exe` utility, and specify the name and location for the CSV file as shown.



```
Administrator: C:\Windows\system32\cmd.exe
E:\Program Files (x86)\VMware\VMware vCenter Site Recovery Manager\bin>dr-ip-customizer.exe -cfg "E:\Program Files (x86)\VMware\VMware vCenter Site Recovery Manager\config\vmware-dr.xml" -csv c:\down.csv -cmd generate_
```

3. Edit the CSV file to provide the IP settings for the virtual machines at the recovery site. The following image shows the edited file for this solution.

A	B	C	D	E	F	G	H	I	J	K	L	
VM ID	VM Name	Adapter ID	MAC Addr	DNS	Dor	Net B	Primar	Se	IP Address	Subnet Mask	Gateway(s)	DNS Server(s)
shadow-vm-36121	TCE-SQL-TPCE1-75	1							172.16.200.75	255.255.255.0	172.16.200.1	172.16.200.75
shadow-vm-36298	TCE-SQL-TPCE2-50	1							172.16.200.75	255.255.255.0	172.16.200.1	172.16.200.75
shadow-vm-36408	TCE-SQL-TPCH1	1							172.16.200.75	255.255.255.0	172.16.200.1	172.16.200.75
shadow-vm-36531	TCE-SQL-TPCH2	1							172.16.200.77	255.255.255.0	172.16.200.1	172.16.200.77

4. Run the utility to upload the new settings to the recovery site vCenter server as shown.



```
E:\Program Files (x86)\VMware\VMware vCenter Site Recovery Manager\bin>dr-ip-customizer.exe -cfg "E:\Program Files (x86)\VMware\VMware vCenter Site Recovery Manager\config\vmware-dr.xml" -csv c:\down.csv -cmd create
```

Note: If you delete or recreate a protection group, you must repeat this process to reapply the IP customizations.

Configuring vCenter SRM recovery plans

Recovery plans are located at the recovery site and define steps for recovering virtual machines. vCenter SRM recovery plans can use the RecoverPoint image access capability to non-disruptively test the failover process. This ensures that the secondary image is consistent and usable.

Testing disaster recovery plans are critical to ensure recovery is reliable. Traditionally, this was a complex, time-consuming, and costly exercise. With vCenter SRM, you can overcome these obstacles by enabling realistic, frequent tests of recovery plans and eliminating common causes of failures during recovery.

By including multiple protection groups in a single recovery plan, all of the associated virtual machines are available to recover as part of that single recovery plan. Figure 42 shows the first step in running the recovery plan.

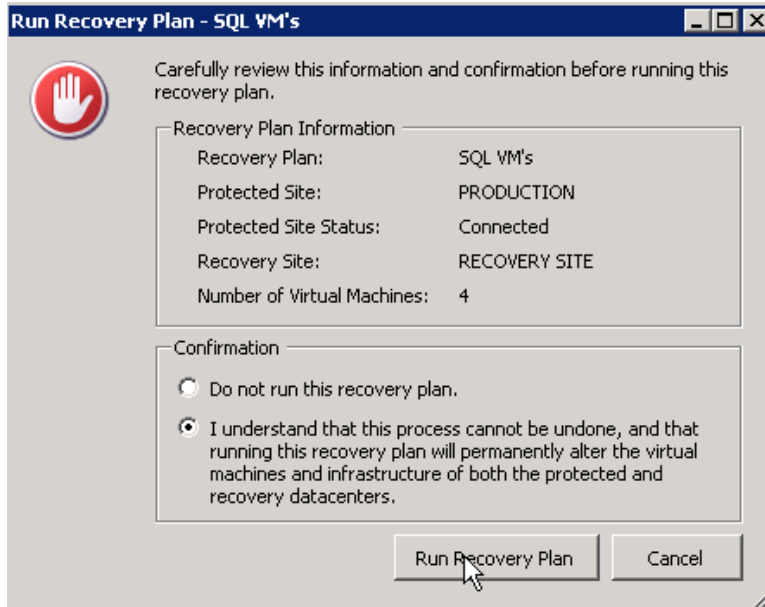


Figure 42. Running the recovery plan

As shown in Figure 43, the prioritized virtual machine is being recovered (restarted) before the other TCE virtual machines in the same recovery plan.

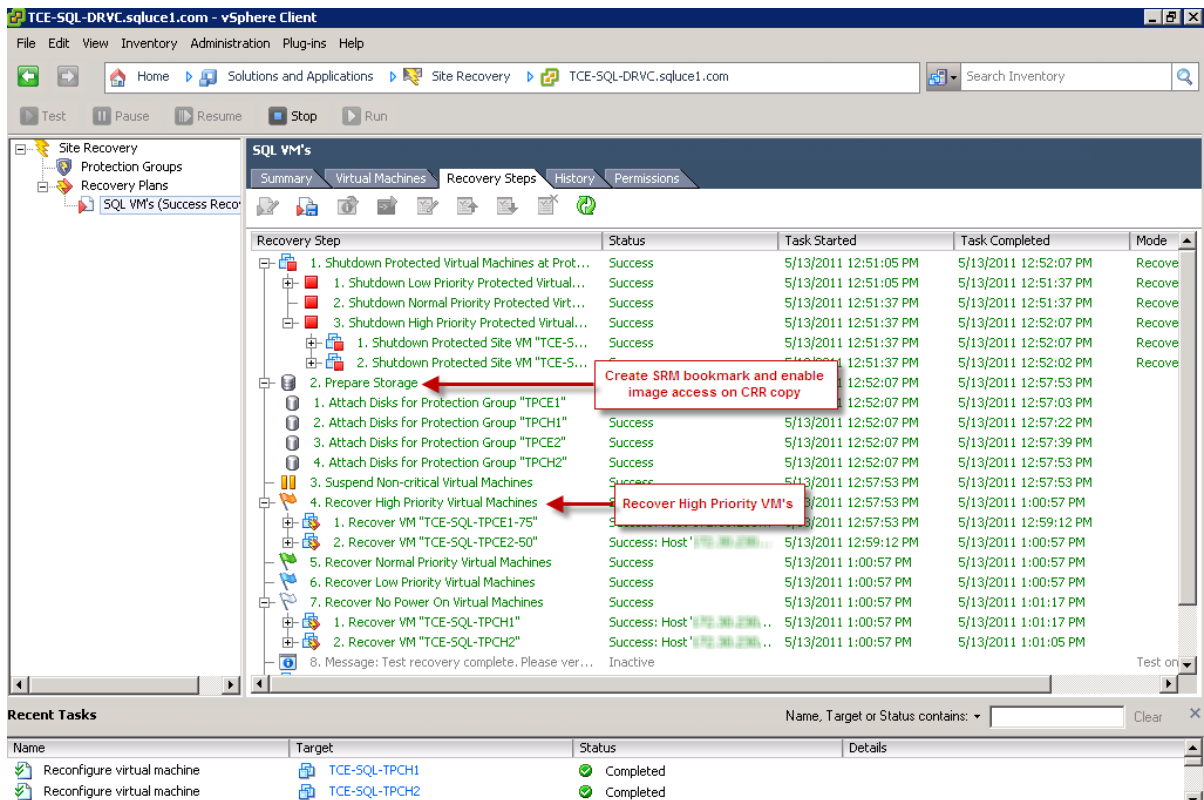


Figure 43. Prioritizing virtual machine recovery

When the failover process is finished, vCenter SRM displays a summary report of the recovery, as shown in Figure 44.

SQL VM's

Description

SQL Server Recovery Group

Start Time: 5/13/2011 12:51:05 PM

Finish Time: 5/13/2011 1:01:17 PM

Total Execution Time: 00:10:12

Mode: Recovery

Overall Result: Success

Figure 44. Failover summary report

Configuring vCenter SRM failover with RecoverPoint CLR

After vCenter SRM successfully completes the recovery plan, and all systems are operational again, you must complete the following manual steps to resume full CRR replication back from the disaster recovery site to production site:

1. Ensure that the group is in maintenance mode and is being managed by RecoverPoint, with SRM only monitoring.
2. Set the RecoverPoint Remote Replica copy as production on the disaster recovery site as shown in Figure 45.

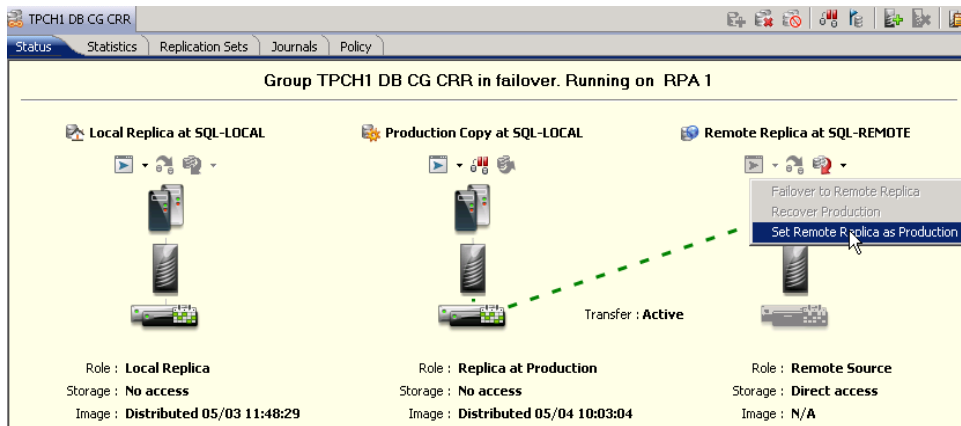


Figure 45. Setting the RecoverPoint remote replica copy as the production copy

3. Because CDP was also on the production site before failover, remove one of the replica data copies on the production site, as shown in Figure 46.

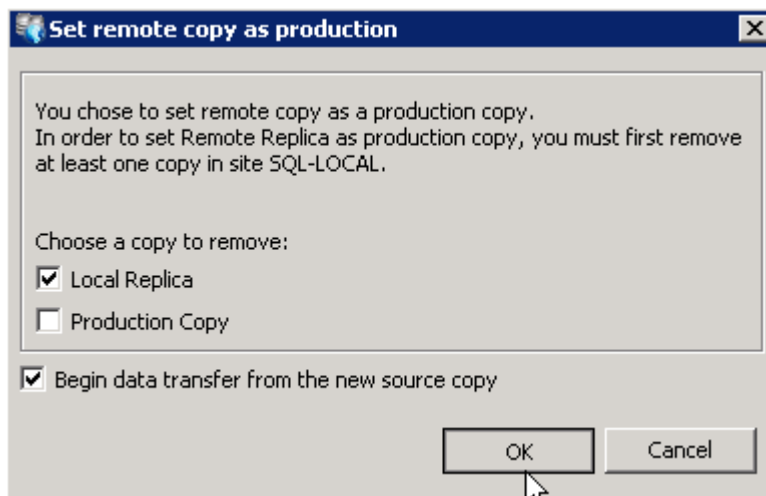


Figure 46. Removing the local replica

After setting the CRR copy as your production copy, you are prompted to choose a copy of the data on the production site to be removed. You must then decide whether to use the production copy or CDP copy as the target for CRR, which is done by removing the unneeded copy, as shown in Figure 47.

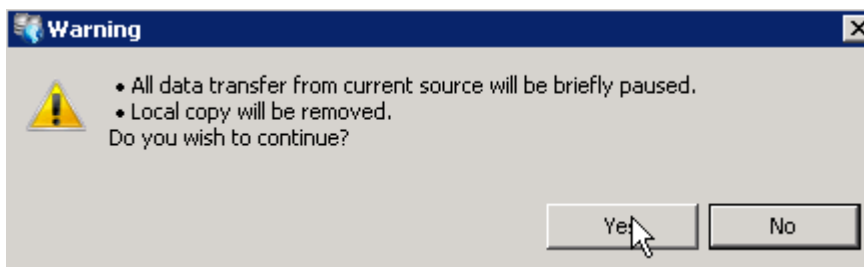


Figure 47. Removing unneeded copy of the data

As part of a RecoverPoint CLR configuration, the production site previously hosted both the production and CDP data copies. This new RecoverPoint replication configuration is CRR, so only one target copy of the data is possible on the production site.

These settings do not affect the recovery of the virtual machines on the disaster recovery site. They are specific to RecoverPoint and are required for configuring new CRR relationships back to the production site.

In the event that a disaster has rendered the production site unreachable, these steps are not necessary until communications with the production site are restored.

If communications with the production site is still available after the recovery, these steps can be scripted in the RecoverPoint CLI. As part of a controlled failover, these commands can be included as a post-script operation in the vCenter SRM recovery plan.

The result of reconfiguring the consistency group is a straightforward RecoverPoint CRR replication from the disaster recovery site to the production site, as shown in Figure 48.

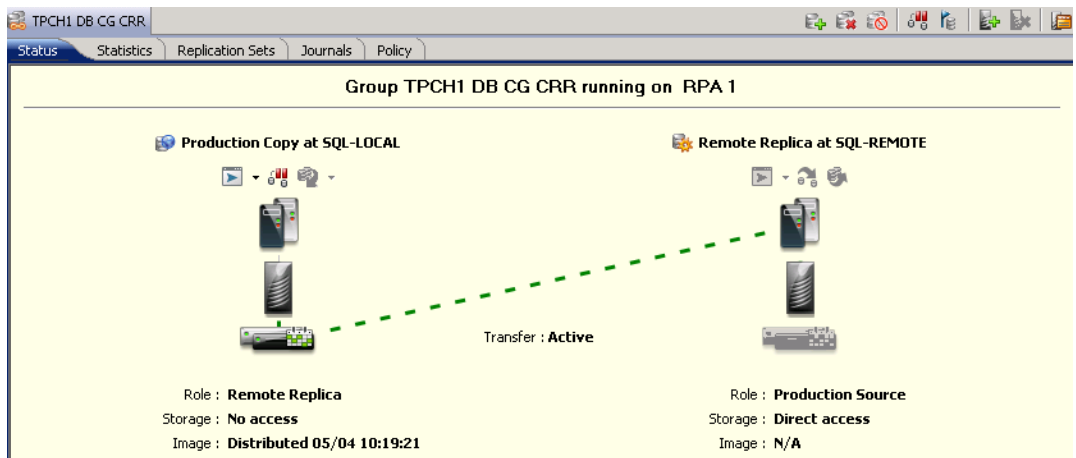


Figure 48. Reconfiguring of consistency group

After a failback completes, the production site resumes production, a full reconfiguration and resynchronization of the CDP copies are required.

Note: It is possible to configure CDP on the disaster recovery site and remain in this configuration, if appropriate.
