iSCSI SAN Solution Design Guide for VSM on the UCS Express Platform

October 2012

The Cisco® IP Video Surveillance (IPVS) solution utilizes the network as a platform to deliver live streaming and on-demand, recorded surveillance video to end users. This design guide introduces the use of Internet Small Computer System Interface (iSCSI) storage area networks (SANs) as a solution to this storage scalability issue in remote office/branch office (ROBO) networks.

Contents

This document includes the following sections:

Americas Headquarters:
Cisco Systems, Inc., 170 West Tasman Drive, San Jose, CA 95134-1706 USA

Introduction

The Cisco IPVS solution utilizes the network as a platform to deliver live streaming and on-demand, recorded surveillance video to end users. The solution is comprised of Video Surveillance Manager (VSM), which is a suite of video management server applications, IP cameras, analog encoders and SAN storage systems.

For ROBO sites that typically consist of fewer than 30 video endpoints and 5 end users, the IPVS solution can suitably be deployed on the Unified Computing System (UCS) Express platform, following the distributed architecture model for IPVS. This architecture consists of a central campus that provides the centralized management framework and one or more “lean” branch sites where the Video Surveillance Manager (VSM) application is deployed.

The UCS Express platform, a part of the Cisco UCS, consists of next-generation, power-optimized, general-purpose x86 64-bit blade servers that are deployed in the Cisco Integrated Services Router (ISR) G2.

The UCS Express platform provides limited internal storage (up to 1 TB) for archived video that is insufficient for meeting the storage requirements in most branch deployment use cases. As a result, alternative scalable storage options should be considered.

The purpose of this design guide is to introduce the use of iSCSI SANs as a solution to this storage scalability issue in ROBO networks. This solution offers significant noted advantages, as well as some caveats discussed further in this guide.

Benefits

The iSCSI SAN solution offers several benefits to the IPVS branch environment, including:

- **Storage scalability**—The internal storage offered by the UCS Express platform is limited to a maximum of 1 TB raw capacity. Utilizing external iSCSI SAN storage provides additional capacity that is only limited by the total available capacity of the attached iSCSI target.
- **Lowers total cost of ownership**—iSCSI leverages existing Ethernet infrastructure as the communication path between initiators and targets. As a result, no additional network hardware is required while providing significant added value.
- **Security**—iSCSI provides for secure peer authentication between initiators and targets to validate the identity of connected nodes. As a result, only permitted hosts are able to access logical unit numbers (LUNs) on the storage target, mitigating data compromise.
- **Block-level storage**—iSCSI allows for a simplified integration to existing block-based SAN arrays and also provides support for formatting the mapped LUNs in any desired native file system format supported by the guest operating system (OS). For VSM, the desired file system for media repositories is XFS.

Caveats

There are a few drawbacks to be considered when deploying iSCSI SAN solutions:

- **"Lossy" Ethernet**—Because iSCSI is deployed over an Ethernet infrastructure, it is prone to packet loss, jitter, latency, and delay. These effects are mitigated by the Transmission Control Protocol (TCP) to a great extent, which has robust and reliable packet delivery mechanisms.

  Note The current version of the Cisco Services Ready Engine (SRE) module only supports 1 GbE iSCSI; the upcoming next-generation UCS-E modules will provide optional support for 10 GbE Fiber Channel over Ethernet (FCOE) iSCSI.

- **Protocol Overhead**—iSCSI protocol data units (PDUs) create additional processing overhead, in addition to the processing of TCP traffic and state machine. Today, this is less of an issue with the advent of faster Intel processors, TCP Offload Engines, and iSCSI Hardware Offload capabilities with iSCSI initiators.

Note Cisco shall not be held responsible for any issues related to third-party storage devices that are used as iSCSI storage targets, including but not limited to hardware and software design and support. Cisco will only provide general baseline requirements, which are noted in this document, that the target device selected needs to satisfy. Support requests for all non-Cisco storage appliances shall not be fielded by Cisco TAC, and must be directed at the respective hardware vendor for further assistance.

Scope

This guide provides best practice recommendations for designing and optimizing iSCSI SAN solutions based on the following products:

- Cisco Services Ready Engine Virtualization (SRE-V) 2.0.1
- Cisco Services Ready Engine 900/910 module
- Cisco Integrated Services Router (ISR) G2
- Cisco Catalyst 3560/3750 Internetwork Operating System (IOS) switches
- Cisco Video Surveillance Manager 6.3.2 MR2

In this guide, the considerations for designing a SAN based on iSCSI are described; specifically, the technology and enablers for iSCSI, and the proposed architecture and design considerations. A description of the key performance metrics and thresholds, and resilience of the solution is also addressed.

Technology Review

iSCSI

Overview

The Small Computer Serial Interface (SCSI) is a collection of standards defining the command sets, signaling, and protocols that have long been used to allow hosts to communicate with storage devices.

The communication medium over which SCSI commands are transported has evolved over time from use of a dedicated parallel SCSI cable attached, back-to-back, between hosts and the storage devices, to today's transport of SCSI over Ethernet (iSCSI), fiber channel (Feature Control Protocol [FCP]), and serial cable (Safety and Security [SAS]).

Architecture

SCSI is a client-server architecture, where the client, referred to as the initiator, issues SCSI tasks, and the server (the target) services the requests. The iSCSI protocol runs over TCP, typically listening on the default, well-known port, TCP 3206.

SCSI commands exchanged between nodes are contained in data structures referred to as command descriptor blocks (CDBs). The CDBs are divided into messages known as iSCSI PDUs, and are transported as TCP payloads to the target and back over established TCP connections.

Each individual input/output (I/O) device on the target is referred to as a logical unit (LU). A target can have multiple such units and, therefore, each is identified by an address—the logical unit number (LUN). Each LUN has an associated queue that is managed by the storage device. The size of this queue (the LUN queue depth) determines the maximum number of I/O requests from the host that can be outstanding at the LUN at any one point in time, which can, in turn, impact performance depending on the level of I/O activity at the host.

All iSCSI nodes have a unique, worldwide identifying name. This identifier allows multiple initiators or targets to either share a common IP address or be accessed by multiple IP addresses. The addressing can either follow the iSCSI qualified name (IQN) or the Institute of Electrical and Electronic Engineers (IEEE) Extended Unique Identifier-64 (EUI-64) format.

After a TCP connection is established, an iSCSI session is formed between the initiator and the target. A host can have multiple sessions to a storage device; for example, when high availability is implemented. The initiator begins the session initiation process by logging on to the target, authenticating (if configured), and exchanging session parameters. Having established the iSCSI session, data transfer can now begin.

Security

The iSCSI protocol provides for security for the session based on the Challenge Handshake Authentication Protocol (CHAP). This authentication between the initiator and target is optional. If enabled, authentication of the initiator by the target would occur before the logon process begins.

With CHAP, an authentication key is shared between the initiator and target. For authentication to be successful, the key must match on both ends.

There are two forms of CHAP:

- **One-way**—The target authenticates the initiator.
- **Mutual**—Both the target and initiator authenticate each other.

Another way to provide security for storage traffic is through use of dedicated virtual LANs (VLANs). This form of security provides Layer-2 separation of different traffic aggregates, allowing only devices that are part of the VLAN to participate.

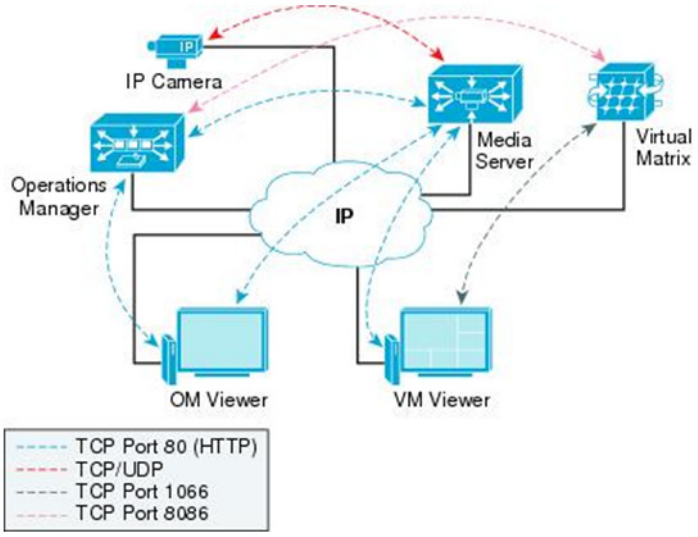
TCP

Overview

The Transmission Control Protocol (TCP) is a unicast, transport-layer protocol in the TCP/IP stack that provides a connection-oriented, full-duplex, reliable, end-to-end, byte-stream service for bulk data and interactive data applications. Essentially, two TCP-speaking applications (in this instance, the iSCSI initiator and target) must first establish a connection before exchanging data, over which data can flow bidirectionally and independently of each other, and that delivery of data segments is guaranteed.

In Figure 1, VSM traffic flows are mostly TCP-based, except in the case of traffic flows from video endpoints, which can also be configured based on Real-Time Transport Protocol (RTP) over User Datagram Protocol (UDP).

Figure 1 VSM Traffic Flow



Understanding how TCP directly works influences the performance and resilience of the iSCSI transport. This, in turn, can have an effect on a VSM's application performance—positively or negatively.

While a comprehensive review of TCP is beyond the scope of this guide, two characteristic mechanisms unique to this protocol are worth briefly examining—flow control and congestion control.

Flow Control

TCP implements flow control by having the receiver notify the sender the amount of data that it is willing to receive; that is, its advertised window (*awnd*). The *awnd* is a measure, in bytes, of the amount of space the receiver has in its buffer to accept incoming data, whose lower limit is the last contiguous, acknowledged sequence number from the sender.

The size of the *awnd* is carried in the 16-bit "Window Size" TCP field, meaning that the maximum TCP window size that can be advertised by this field is 65,536 bytes. However, this value can be increased by use of window scaling, mostly for long fat networks (LFNs), or satellite links. By default, the window size in the ISR G2 is 4,128 bytes without window scaling and 65,536 bytes, if only one of the neighbors is configured for window scaling.

As data is acknowledged and buffer space is freed up at the receiver, the window advances ("slides forward"), allowing more data to be received from the sender. This phenomenon is referred to as "sliding window". The sender can transmit up to the *awnd* size, unless it receives a "window update" indicating an increase or decrease in window size. When *awnd* is zero, the sender stops sending data; however, it continuously probes the receiver for a window update.

Essentially, TCP flow control mechanisms, such as sliding window, provide explicit information about the receiver's state back to the sender so that it can adaptively regulate its transmission rate to prevent overflow of the receiver's buffer with incoming data.

If video data is being streamed to a viewing client from the VSM Media Server, the client is unable to render the images fast enough onscreen, causing the input buffers to fill up. The TCP sender (the server) will, therefore, be notified (via window update) by the TCP process at the receiver. The TCP window update could indicate a smaller window size or a window size of zero, which indicates that it cannot receive more data.

Congestion Control

Congestion in a network occurs when intermediate nodes receive more data than they can buffer and transmit, resulting in performance degradation due to excessive delay and packet loss.

Once a connection is established, TCP adopts a conservative approach to sending data by initially transmitting modest amounts of packets, based on its sending window, referred to as the congestion window (*cwnd*). The initial value of *cwnd* is typically set equal to the sender's maximum segment size (MSS), which, by default, is 536 bytes.

The *cwnd* increases exponentially by the value of the sender's MSS for every received acknowledgement from the receiver, invoking the slow-start algorithm as it probes the network to determine available capacity to handle data, as well as determine a value for the slow-start threshold. This exponential increase continues until the *awnd* is exceeded, the slow-start threshold (*ssthresh*) is exceeded, or the network's capacity is exceeded, as indicated by packet loss.

TCP detects packet loss in one of two ways:

- **Expiration of the Retransmission Timeout (RTO)**—Each TCP segment sent with the Timestamps option has a timer that measures how long it takes to receive an acknowledgment code (ACK) from the receiver once sent. If the timer, based on the smooth round-trip time estimate, expires before the ACK is received by the sender, it is assumed that the packet has been lost on the network.
- **Receipt of duplicate acknowledgements**—When TCP receives an out-of-order segment (a segment with a non-contiguous sequence number), a duplicate ACK (segment with the same ACK field value as the last) is forwarded to the sender for that packet and all successive packets received, inferring packet loss or reordering.

TCP recovers from packet loss either by:

- **Timer-based retransmission**—If no ACK is received after the RTO, the sender retransmits the last packet sent. This algorithm introduces additional delay because the sender must wait until the RTO expires.
- **Fast retransmit**—After three duplicate ACKs (*dupthresh*) are received by the sender indicating packet loss, the sender immediately retransmits the packet inferred by the duplicate ACK without waiting for the RTO to expire.

When network loss is detected—either because the receiver or the network is too slow—the TCP sender is required to slow down its transmission rate in response to the congestion.

During normal operations, one of two TCP algorithms is running that prescribes the data flow on the network between the sender and receiver:

- **Slow-start mode** is invoked either when a connection is initially established or when packet loss is detected due to RTO expiration. When in slow-start mode, the *cwnd* is set to 1 MSS, and the slow-start threshold (*ssthresh*) is set to half the size of the current unacknowledged data, i.e. the flight-size. The transmission rate increases exponentially as before until the slow-start threshold, then switches to a more conservative linear growth progression phase referred to as *congestion avoidance*.
- **Congestion avoidance mode** is invoked when the *cwnd* is greater than ssthresh. In this phase, the amount of sent data grows much slower and more cautiously than in slow-start mode, in a bid to balance the prospect of gaining more capacity to send data while avoiding opening the *cwnd* too large too fast, thereby increasing the probability of congestion and subsequent packet loss.

The reliable transmission and delivery mechanisms built into TCP are useful to Cisco IPVS environments, because they help ensure that video transport to end users is resilient and can recover from network-related issues, such as packet loss, latency, and jitter. However, if these issues occur often and consistently, the performance of VSM and quality of the user experience may be significantly degraded and, as such, must be promptly addressed.

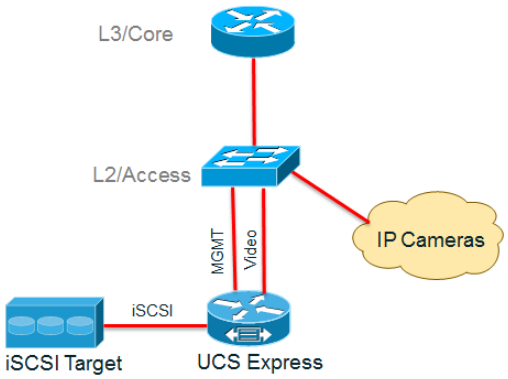
Design Considerations

This section describes the recommended architecture and best practice considerations for designing iSCSI SAN solutions in Cisco IPVS branch environments.

High-Level Design

At a high-level, Figure 2 illustrates the proposed lean branch architecture.

Figure 2 Lean Branch Architecture



This topology represents a typical, small branch environment, where the L3/Core router functions as the default gateway and facilitates inter-VLAN routing ("router-on-a-stick" configuration). The L2/Access gigabit switch provides connectivity to the LAN environment, including the IP cameras deployed around the branch location, as well as the UCS Express platform running on an ISR G2.

Management, IP video, and storage traffic are all separated at Layer 2. The iSCSI target connects directly to the ISR G2 via the Enhanced High-Speed Wide-Area Network (WAN) Interface Card (HWIC) module.

The following sections describe the recommended configuration and best practices for the proposed design.

Compute Platform

The compute platform consists of the baseline-recommended configuration, outlined in this section.

Cisco ISR G2

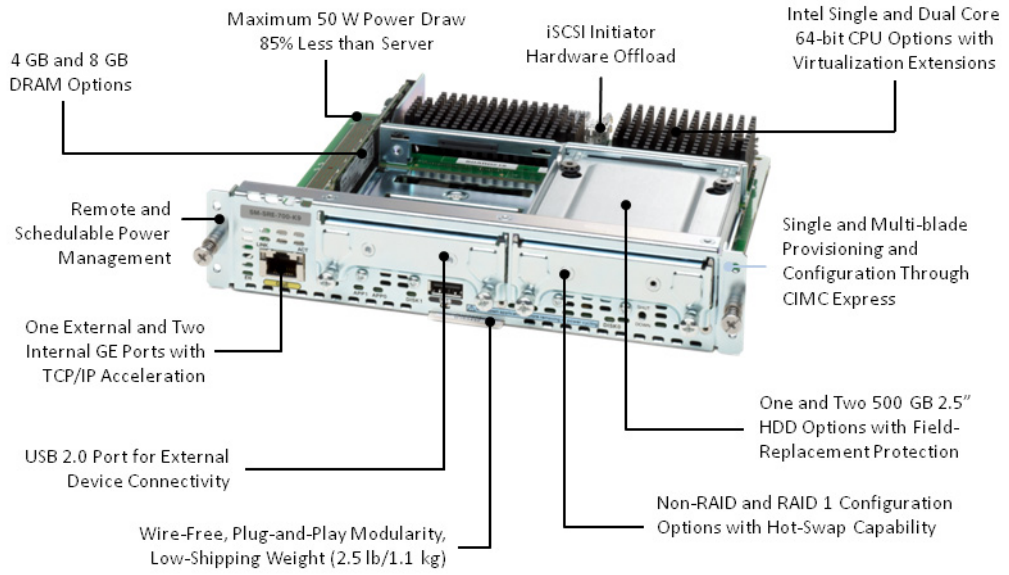
The Cisco ISR G2 provides a highly secure and reliable platform for scalable multiservice integration at enterprise and commercial branch offices of all sizes and small-to-medium-sized businesses. The ISR G2 runs a single, universal IOS software image on a multicore network processor providing up to 350 Mbps WAN performance with services.

Note For the IPVS branch environment, we recommend the 2911, 2921, 2951, or 39xx ISR G2 models.

SRE Module

The SRE is a Cisco ISR G2 router service module that provides the functionality of a compact, multipurpose x86 blade server (see Figure 3).

Figure 3 Cisco SRE x86 Blade Server



Each module has its own CPU, memory, storage, and network interfaces that operate independently of the host router.

The module consists of three network interfaces:

- An internal Broadcom Peripheral Component Interconnect Express (PCIe) gigabit interface with TCP Offload Engine and iSCSI hardware offload that interfaces with the SM1/0 router interface
- An internal Broadcom gigabit interface with TCP Offload Engine and iSCSI hardware offload that interfaces with the router's Multigigabit Fabric (MGF) interface via SM1/1
- An external Intel gigabit Ethernet interface

There are two memory options to choose from with the SRE—either 4 GB or 8 GB random-access memory (RAM). By default, the amount of memory provisioned to the VSM virtual machine (VM) in the VM template provided by Cisco, is set to 3.5 GB. This value has been validated to work well for 4 GB RAM installations. If the 8 GB RAM option is selected, then the memory allocation can also be increased by editing the VM settings through the vSphere client.

Note Cisco recommends that changes to the VM's hardware profile should only be made when powered off. For more information on configuring the VM settings, see the VMware documentation at http://pubs.vmware.com.

Note The configuration maximums must still adhere to the stipulated settings in the VSM on the UCS Express datasheet.

SRE-V

Cisco Services Ready Engine Virtualization (SRE-V) is the hardware virtualization platform that is powered by the enterprise-class VMware ESXi bare-metal hypervisor. SRE-V is specifically optimized for SRE blades and tailored for branch deployment use cases.

Note The SRE-V 2.0.1 hypervisor enables the VMware ESXi 5.0 release to be provisioned on the SRE. Several of the features that are a part of the ESXi 5.0 release are included in this customized release. However, some features, while may be available, are not supported by the Technical Assistance Center (TAC), such as vMotion and the Dynamic Resource Scheduler (DRS).

Enhanced High-Speed WAN Interface Card

The Enhanced High-Speed WAN Interface Card (EHWIC) modules are low-density, Gigabit Ethernet integrated switch modules that offer small-to-medium-sized businesses and enterprise branch-office customers a combination of switching and routing integrated into a single device.

They offer line-rate Layer-2 switching across Gigabit Ethernet ports, with up to 20 W enhanced Power over Ethernet (ePoE) options on all 4 or 8 ports. Each port has its individual quality of service queues and the MGF is enabled for direct, module-to-module communication.

Note We recommend using EHWIC to provide Layer-2 traffic separation between video and storage traffic aggregates, while taking advantage of line-rate switching.

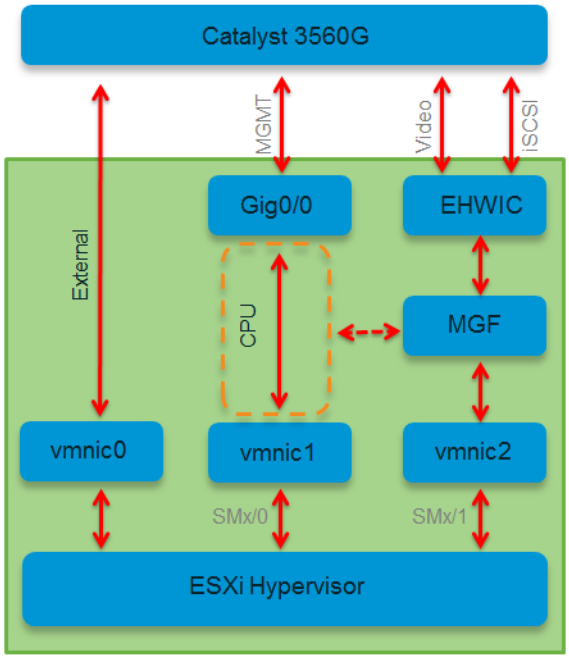
Network Platform

Network Interfaces

Figure 4 displays an overview of the network hardware platform. Three interfaces are exposed to the hypervisor—**vmnic0**, **vmnic1**, and **vmnic2**.

Note **vmnic0**, which corresponds to the external interface, is only visible once the Ethernet connectivity to an upstream network node has been enabled in the UP state. However, this interface is not visible through the IOS command-line interface; therefore, routing cannot be configured on this interface. In other words, the external interface behaves not as a routed interface, but as a switched interface.

Figure 4 Network Hardware Platform Overview



All in-band management traffic to the SRE module, such as Telnet and Secure Shell (SSH) is routed through SM1/0 (presented as **vmnic1** in the vSphere client). Cisco Integrated Management Controller Express) traffic is routed through the embedded service engine interface.

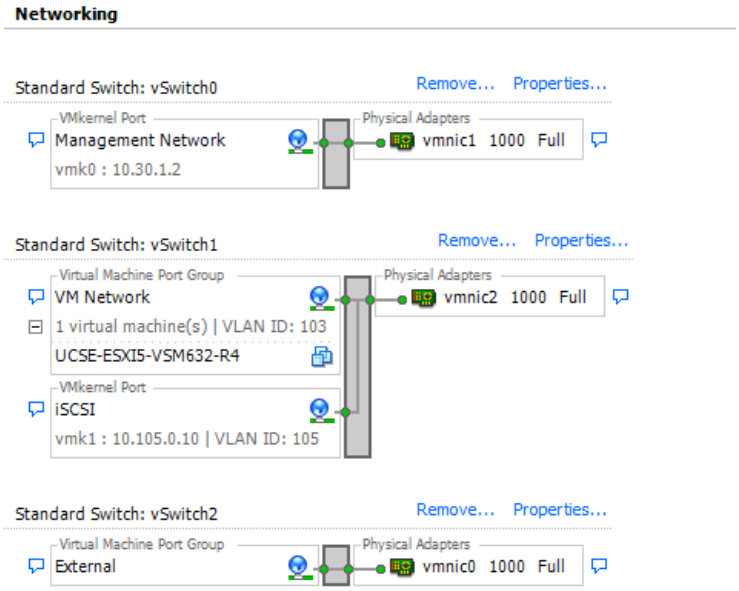
To achieve high performance for iSCSI over the network, a gigabit switch fabric is required, which means that end-to-end links between the SRE and the iSCSI target should be full-duplex, gigabit Ethernet links.

Note We recommend to directly connect the iSCSI SAN device to the EHWIC module on the ISR. This enables two key benefits:

- **Use of the MGF interface**—The MGF allows for network traffic to flow directly between the Service Module (SM) and HWIC module without going through the CPU. Fewer CPU interrupts are created, thus lowering the total CPU utilization.
- **Layer-2 traffic separation**—Because the SMx/1 interface is a switch port, it is configured as a trunk, which allows the creation of switch virtual interfaces (SVIs) for different VLANs. This enables IP video from endpoints to be separated from SAN traffic.

Note We recommend to configure iSCSI only on the **vmnic2** interface. While both **vmnic1** and **vmnic2** both support TCP Offload Engine and iSCSI hardware offload, **vmnic1** is designed to be specifically for management traffic to the ISR and is throttled; see Figure 5.

Figure 5 Switch Port Traffic



The **vmnic2** interface is a switch port configured as a trunk that allows for creating multiple VLANs and SVIs for flexible routing of traffic from the SRE.

Rapid Spanning Tree Protocol

We recommend configuring the Rapid Spanning Tree Protocol (RSTP) and enabling Portfast optimization on the switch ports, on the EHWIC on to which the iSCSI target device is connected.

Portfast allows for an access port to be immediately placed into the forwarding state once the port is administratively enabled and is physically functional. By skipping the listening and learning stages, Spanning Tree Protocol (STP) convergence time can be reduced to less than 5 seconds, compared to 30 seconds with default forwarding delay timers.

Jumbo Frames

We recommend configuring jumbo frames on all network nodes in the path of the storage traffic. Jumbo frames are Ethernet frames larger than the default of 1,500 bytes and are typically 9,000 bytes in size. By increasing the payload, more data traffic can be transported in much fewer frames, thus increasing throughput. Actual results depend on the specific environment; therefore, we recommend testing to gauge the expected performance impact.

Note The Maximum Transmission Unit (MTU) of both the vSwitches and vmkernel interfaces within the ESXi host also must be configured for jumbo frames; otherwise, fragmentation occurs and the tangible benefits is are not realized.

Ethernet Flow Control (IEEE 802.3x)

Ethernet implements flow control at Layer 2 based on the 802.3x standard specifications. Due to the increased bandwidth and network throughput, switches can begin to experience congestion if they cannot forward out frames at a fast-enough pace to keep up with incoming data flows.

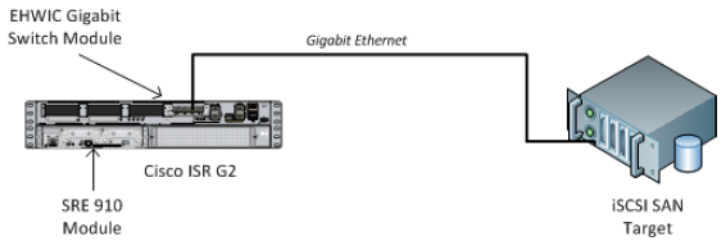
To mitigate packet loss due to tail-drop of incoming flows as a result of full buffers, 802.3x specifies a hop-by-hop, flow-control scheme for full-duplex Ethernet networks. Ethernet flow control works by monitoring the data size in the input buffers, and if it crosses a predefined threshold, the switch is prompted to send a PAUSE frame to stop data transmission from the source. Once the congestion situation is alleviated, the switch sends a transmit off (XOFF) frame that signals the resumption of transmission.

Note In a homogenous, gigabit-switch fabric environment, Ethernet flow control actually can help TCP flow control when handling congestion, but in mixed 10/100/1000-Mbps fabric environments could cause head-of-line blocking, which significantly reduces TCP throughput and degrades performance. Therefore, we recommend (for gigabit switch fabrics) that Ethernet flow control should be enabled in receive-only mode, where the integrated switch only responds to congestion situations.

Storage Platform

In general, the storage infrastructure and connectivity between hosts and storage devices adopts the model displayed in Figure 6.

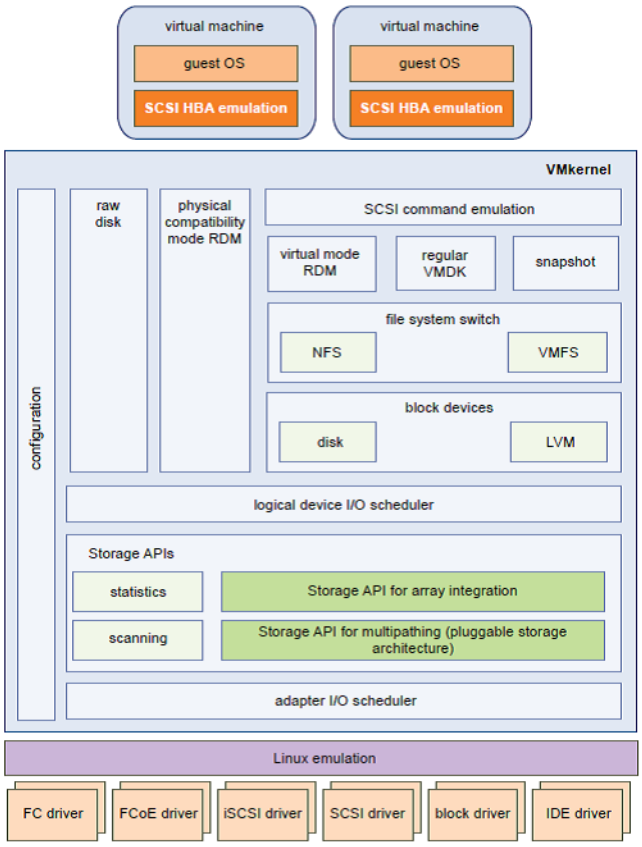
Figure 6 Storage Infrastructure and Connectivity



Note In the case of the SRE, only iSCSI SANs are currently supported.

The storage platform is comprised of both the storage stack within the vmkernel and the iSCSI target device. The vmkernel storage subsystem manages the host bus adapter (HBA), datastores, and device emulation, as shown in Figure 7.

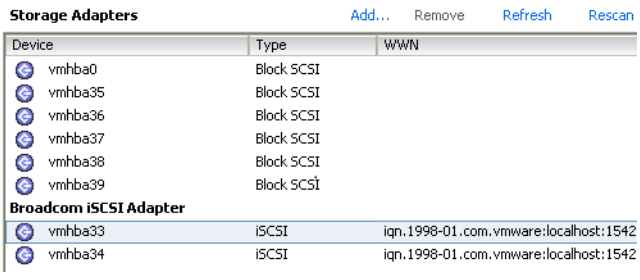
Figure 7 Storage Platform—VMkernel and Storage (www.vmware.com)



The HBA's purpose is to facilitate the session establishment between initiators on the host side and the storage processors on the array. On the SRE, Figure 8 illustrates that the iSCSI initiator could either be a:

- **Software adapter**—The host establishes connectivity to the storage array through standard network adapters. If this adapter is used, Cisco recommends that the vmkernel interface is bound to **vmnic2**.
- **Hardware-dependent adapter**—The host connects through dedicated iSCSI hardware HBA, but still relies on VMware networking for IP addressing and routing. The SRE has a dedicated Broadcom iSCSI adapter with a TCP Offload Engine enabled and iSCSI hardware offload and, as such, is the recommended adapter to use for configuring iSCSI connectivity.

Figure 8 iSCSI Adapter Options



Once the iSCSI session has been established between the ESXi host (initiator) and the storage array (target), the volumes that have been mapped to LUNs on the storage device can now be configured within the vSphere client as available datastores for the system.

Datastores

A datastore, in relation to vSphere, refers to a logical storage unit or container that abstracts the functional and operational details of underlying storage devices, thereby providing a uniform model for storing VM files and images. By definition, a file system refers to a hierarchy of directories used to organize files in a single disk or partition.

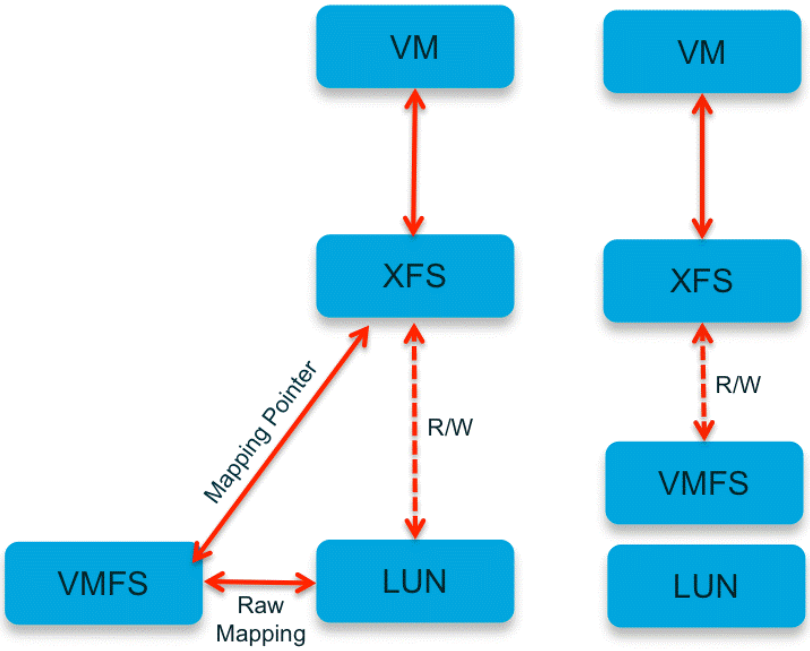
VMware supports two datastore file systems:

- **Network File System (NFS)**—A network protocol backed by a high-performance file system that is optimized for file sharing, and is used to access network attached storage (NAS) devices. The shares are exported from the NAS device and mounted within the hypervisor.
- **Virtual Machine File System (VMFS)** —A high-performance, clustered file system optimized for VMs and leverages shared storage to allow concurrent access to the same pool of shared storage by multiple ESXi hosts.

From the VM perspective, each virtual disk appears as if it were a physical disk connected to a SCSI controller. When either NFS or VMFS datastores are provisioned to a VM as a virtual hard disk, a virtual machine disk (vmdk) file is created. The maximum size of a vmdk today is 2 TB minus 512 B and is presented as a block device; that is, it can be formatted within the guest OS with traditional native file systems, such as XFS, ReiserFS, ext3, and so on.

To obtain direct access to a LUN, Raw Device Mappings (RDMs) must be configured. An RDM is a mapping file that resides on a VMFS volume and acts as a proxy for a raw disk (see Figure 9). The mapping file contains metadata used to manage and redirect access to the physical device.

Figure 9 Raw Device Mapping and VMFS Datastores



Note Currently, only RDM and VMFS datastores are supported for the IPVS branch environment. NFS datastores are not supported. Further, Cisco recommends that VMFS datastores should only be considered for disk arrays of less than 10 TB usable capacity. To provide support for disk arrays greater than 10 TB, Cisco strongly recommends the use of RDM datastores.

iSCSI Target

The reliability and performance of the iSCSI target ensures that the experience quality of the IP Video Surveillance solution is maintained at a high level. Therefore, Cisco recommends that the iSCSI target selected is an enterprise-class storage device for the Branch, which meets or exceeds the following baseline requirements:

- Hardware specifications
  - 1 Gigabit Ethernet iSCSI controller adapter
  - 1 GB battery-backed controller write cache
  - Serial Advanced Technology Attachment (SATA) 720 revolutions per minute (RPM)hard disk drives
  - Redundant Array of Independent Disks (RAID) 5 or 6 array configuration
  - Redundant power supplies
- Throughput specifications
  - Maximum sequential write input/output (IO): 100 MBps
- IOPS specifications
  - Minimum input/output operations per second (IOPS): 300

The iSCSI target is required to sufficiently handle IPVS workloads based on the following requirements:

- One (1) VSM 6.3.2 VM
- Maximum of 32 video streams
- No audio streams
- Maximum combined video throughput from cameras is 60 Mbps

Application Platform

Note It is essential to ensure that the performance of the VSM application is not adversely affected due to the overprovisioning the number of endpoints or oversubscription of system resources. One way to avoid these scenarios is to adhere to the configuration maximums, as stipulated in the VSM UCS Express datasheet http://www.cisco.com/en/US/prod/collateral/vpndevc/ps6918/ps9145/ps9152/data_sheet_c78-703863.html.

Note In addition to following the design guidelines noted, it is also considered best practice to periodically examine the VSM logs to ensure that any errors or warnings that may be reported are attended to before the application performance and usability is adversely affected.

The application logs, described in this section, are relevant in assessing performance.

IMS Log

The IMS log is the Main Server log and is used to record errors, warnings, and notifications relating to Media Servers, video endpoints (i.e., IP cameras, analog encoders, and inbound streaming information).

The IMS log can be located at:

```
shell> more /usr/BWhttpd/logs/ims.log
```

Any RTP packet drops (logged as `Missed <x> RTP packet(s))` as a result of lost packets on the network, from the endpoint device to the server, is recorded here. RTP detects packet loss based on reception of sequence numbers that are out of order.

Note To ensure optimal application performance, this log should indicate no RTP errors, or at most, very few over long time periods with no consistent or repeated patterns.

MEDIAOUT Log

The MEDIAOUT log records errors, warnings, and notifications related to outbound streaming requests and media sessions from the Media Server. This log indicates whether end user clients are keeping up with the outbound Media Server streams or if input buffers are filling up and packets are being dropped.

The MEDIAOUT log can be located at:

```
shell> more /usr/BWhttpd/logs/mediaout.log
```

Whenever latency is detected in sending out video data between the media server and the viewing client, an error message is logged as `Proxy <p_xx> writing to client <ip_address> took <yy> msecs.`

Note If archived video from the iSCSI SAN is the content being requested by the end user, the issue is likely not with the iSCSI access, rather with the client to server connection. Cisco recommends to ensure that the MEDIAOUT log does not indicate any latency so as to provide a higher level guarantee on the performance of the VSM application.

XVCRMAN Log

The XVCRMAN log records all errors, warning, and notifications relating to outbound streaming traffic to storage devices; in this case, the iSCSI SAN target. This log provides a status of the iSCSI SAN's health as it relates to the transport of video data and command I/O between the Media Server and the iSCSI target.

The XVCRMAN log can be found at:

```
shell> /usr/BWhttpd/logs/xvcrman.log
```

Recording issues noted by the XVCRMAN log could include `Unable to read frame for archive a_p_<xx>`, which would indicate that the Media Server cannot successfully request and retrieve the on-disk video data for the specified archive due to a storage connection failure.

In addition, messages logged as `Queue overflow, capture thread appears failed` indicate that the application is unable to keep up recording incoming video data. These errors indicate video data was not recorded.

Note Cisco highly recommends verifying that no recording errors are recorded, and if so, that they are addressed expeditiously to mitigate video loss.

Performance Considerations

The performance of the ESXi host (in this case, the SRE) must be carefully considered to ensure that VSM performs as expected based on its application configuration. As a starting point, the VSM server is required to adhere to the configuration guidelines specified in the VSM for UCS Express datasheet at http://www.cisco.com/en/US/prod/collateral/vpndevc/ps6918/ps9145/ps9152/data_sheet_c78-703863.html.

The performance indicators discussed in the following sections are derived from esxtop output generated by the host. For more information on collecting, analyzing and interpreting esxtop output, consult VMware documentation at http://kb.vmware.com/kb/1008205.

CPU

Two main metrics to consider, relating to CPU performance, are CPU utilization and CPU ready state.

CPU Utilization

In this context, %USED refers to the percentage of CPU resources that are being used by the VSM VM. When the CPU utilization is high, it is an indication that the CPU resources are constrained and possibly oversubscribed.

Note Cisco recommends that the observed average utilization should not exceed 80%. If it does, this means that CPU resources are stretched and VSM performance may be impacted. Examining the list of processes running may provide insight into which threads are consuming the most resources.

CPU Ready State

The %RDY metric measures the percentage of time when the VM was ready to run, i.e. in the ready state, but no CPU resources were available to be scheduled to execute. Whenever CPU oversubscription is suspected, this metric should be observed to confirm its occurrence.

Note Cisco recommends that the ready time should not exceed 20%.

Memory

Active Memory

The %ACTIVE metric measures the percentage of memory that the vmkernel believes is being actively used by the VM over time. This metric is taken as a percentage of total machine memory. If this metric is high, or is increasingly approaching the provisioned virtual memory, it is indicative of increased activity and load within the VM, or the memory is underprovisioned.

Note Cisco recommends that the active memory does not exceed 80% (~3 GB).

Memory Swapping

Swapping involves saving the state of VM memory to disk swap files as a last resort to ESXi to reclaim unused memory. In and of itself, this occurrence is not significantly detrimental; however, high swap-in/swap-out rates are. If it takes an inordinately long time period to execute read and write I/O commands, it could be a clear indication of memory oversubscription. Either, the amount of VM activity must be reduced (fewer or lower bit-rate streams and archives), or memory must be increased.

Note Ideally, Cisco recommends that no swapping should occur; however, if it does occur temporarily, the rate of page swapping should NOT be high. This threshold varies and should be considered on a case-by-case basis.

Disk

When gauging the performance of a storage array, disk latency is arguably the most important metric to monitor. There are three measurements that comprise overall latency:

- **Device latency**—The average amount of time a storage system takes to service a single request, from the HBA to the platter.

Note Cisco recommends that the average latency should not exceed 50 ms/command.

- **Kernel latency**—The average amount of time the vmkernel stack takes to process a SCSI command. We recommend that the average latency should not exceed 10 ms/command.
- **Queue latency**—The average amount of time a SCSI commands spends in the HBA driver queue; also forms a part of the kernel latency.

Note Cisco recommends that the average latency should not exceed 10 ms/command. If these average latencies are consistently above the stipulated thresholds over time, appropriate steps should be taken for fault isolation to ensure that the optimal performance of VSM is not adversely affected.

Network

The outbound throughput of iSCSI traffic can be measured by observing the egress traffic rate of the vmkernel interface that is associated with the iSCSI adapter. Typically, a separate vmkernel interface would be created to carry iSCSI traffic, distinguishing from the vmkernel interface that carries management traffic through the ISR. (The average throughput observed is expected to vary based on the stream settings selected.)

Note iSCSI traffic generally has a "bursty" profile in nature, with maximum bursts of about 600 Mbps observed. Therefore, ensure that a gigabit switch fabric is implemented end-to-end and that Ethernet switches along the path do not flag this traffic profile as a unicast packet storm.

Resilience Considerations

The ability to provide consistency and predictability of VSM operations in the event of iSCSI SAN failure, and to quickly recover in the event of transient or permanent unavailability of the iSCSI target, is critical to ensuring that the solution is highly resilient.

The following sections consider three relevant scenarios that depict expected behaviors when the iSCSI connection fails from the perspective of the Media Server. The failure scenarios are based on power failure simulations, although it could also be due to a network failure. In both instances, the hypervisor detects similar conditions—unavailability of the storage device.

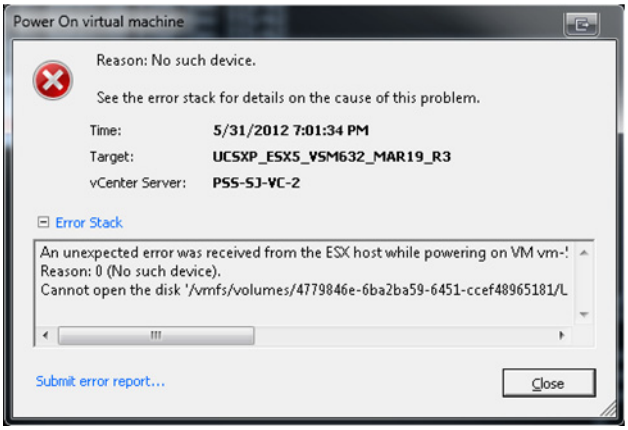
Scenario 1

The first use case describes the condition when the UCS Express platform; that is, the ISR G2, the SRE and SRE-V, and the iSCSI SAN target are both in an initial powered down state. If the UCS Express platform is brought online, while the iSCSI SAN target is still offline, the VSM VM cannot be powered on.

The option to manually power on the VM is dimmed (offline). This is also true for VMs that are set to automatically power on with the host; the VM remains in an offline state.

In the vSphere client, an error message is returned; see Figure 10.

Figure 10 VM Error Message

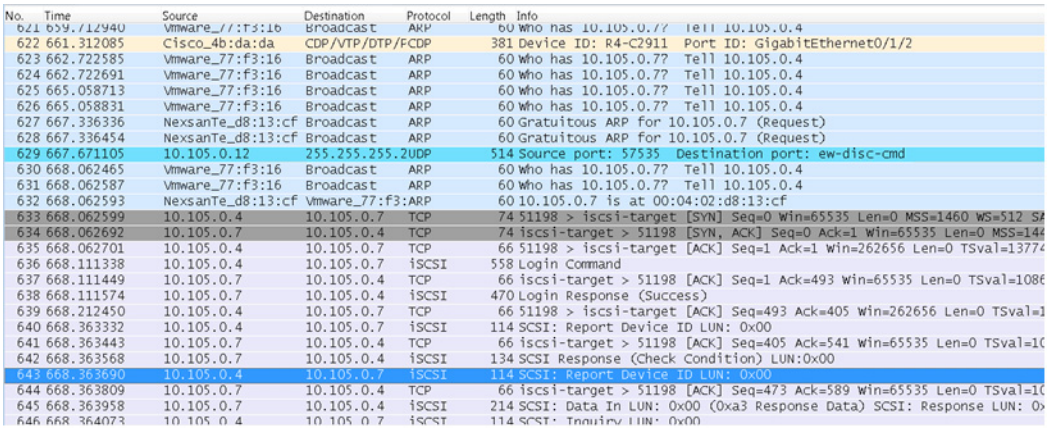


In Figure 10, the VM cannot be powered on due to the missing virtual hard disk that resides on the iSCSI SAN target. This condition is known as All-Paths-Down (APD) because the host is unable to contact the storage array on all paths.

In an APD situation, the hypervisor continues to send I/O from the hosted management agent until a response is received from the storage array. However, all TCP packets from the guest OS, including VSM I/O, are retried for up to a maximum of 15 times before the connection is closed. This duration could last between 13–30 minutes based on the RTO value and the exponential backoff algorithm.

If the storage array is subsequently powered on and brought online, and STP Portfast is enabled on the switch port that the iSCSI target is connected to on the EHWIC, the link physical state and line protocol enter the UP state and iSCSI traffic begins to flow almost instantly. Figure 11 details the sequence of events.

Figure 11 Scenario 1—Sequence of Events



At line 623, the hypervisor is still attempting to contact the storage array by sending Address Resolution Protocol (ARP) broadcasts on the segment, to try and resolve the Media Access Control (MAC) address of the iSCSI target IP configured in the initiator settings.

At line 627, the storage array's switch port is in the UP state and the device then announces its presence by sending out a gratuitous ARP message to inform other network nodes to update their address tables.

At line 632, the storage array responds to the ARP requests from the hypervisor and then subsequently begins the TCP three-way handshake.

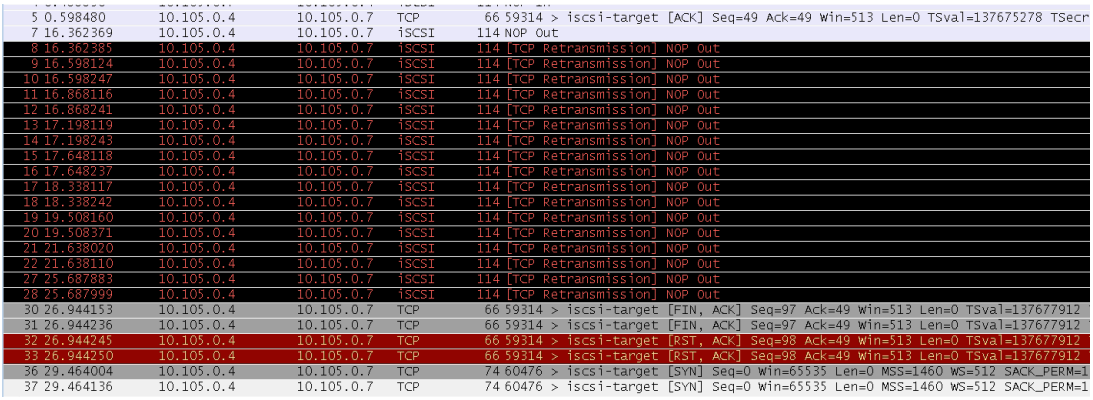
After successful connection establishment, the upper layers begin to exchange traffic.

Scenario 2

This use case considers the state when the VSM VM is in the process of booting up, and the iSCSI SAN target is online and reachable.

If the iSCSI SAN target is powered off during the VM boot up process, the VM is immediately halted. To the VM, a constituent storage device has been ungracefully removed, and from the hypervisor's perspective, the respective LUN has been incorrectly misrepresented without following proper unmounting procedures. This condition results in an APD state. Figure 12 details this sequence of events.

Figure 12 Scenario 2—Sequence of Events



5	0.598480	10.105.0.4	10.105.0.7	TCP	66 59314 > iscsi-target [ACK] Seq=49 Ack=49 Win=513 Len=0 TSval=137675278 TSecr
7	16.362369	10.105.0.4	10.105.0.7	iSCSI	114 NOP Out
8	16.362385	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
9	16.598124	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
10	16.598247	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
11	16.868116	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
12	16.868241	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
13	17.198119	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
14	17.198243	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
15	17.648118	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
16	17.648237	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
17	18.338117	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
18	18.338242	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
19	19.508160	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
20	19.508371	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
21	21.638020	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
22	21.638110	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
27	25.687883	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
28	25.687999	10.105.0.4	10.105.0.7	iSCSI	114 [TCP Retransmission] NOP Out
30	26.944153	10.105.0.4	10.105.0.7	TCP	66 59314 > iscsi-target [FIN, ACK] Seq=97 Ack=49 Win=513 Len=0 TSval=137677912
31	26.944236	10.105.0.4	10.105.0.7	TCP	66 59314 > iscsi-target [FIN, ACK] Seq=97 Ack=49 Win=513 Len=0 TSval=137677912
32	26.944245	10.105.0.4	10.105.0.7	TCP	66 59314 > iscsi-target [RST, ACK] Seq=98 Ack=49 Win=513 Len=0 TSval=137677912
33	26.944250	10.105.0.4	10.105.0.7	TCP	66 59314 > iscsi-target [RST, ACK] Seq=98 Ack=49 Win=513 Len=0 TSval=137677912
36	29.464004	10.105.0.4	10.105.0.7	TCP	74 60476 > iscsi-target [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=512 SACK_PERM=1
37	29.464136	10.105.0.4	10.105.0.7	TCP	74 60476 > iscsi-target [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=512 SACK_PERM=1

When the storage array is taken offline, the hypervisor's TCP stack attempts to retransmit the packet because the RTO expires without receiving any acknowledgements for data transmitted. The hypervisor then closes the session by issuing TCP resets to quickly fail the connection and notify lower layers of the iSCSI SAN target connectivity issue.

At this point, the hypervisor falls back to sending ARP broadcasts on the segment indefinitely to locate the missing storage array, as discussed in Scenario 1, page 20.

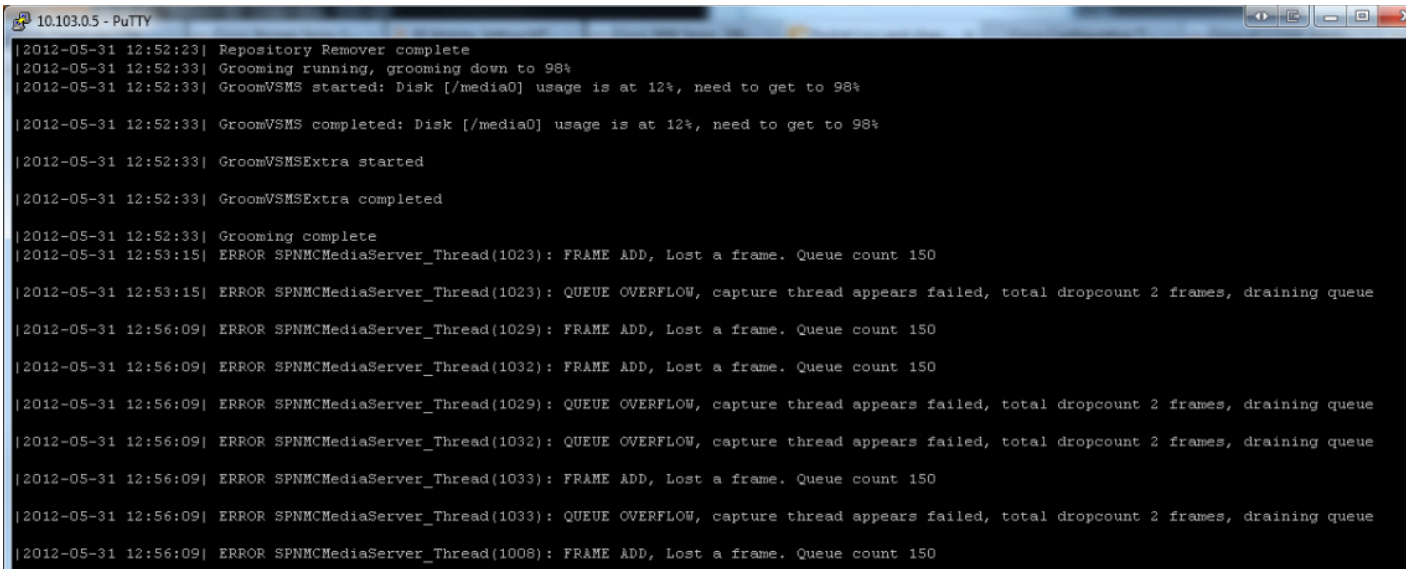
Scenario 3

This use case considers the behavior of the VSM application when an iSCSI SAN target is powered off and becomes permanently unavailable.

While during normal operations (VSM services are running an fully functional and the storage array is available) the iSCSI target is subsequently taken offline, any archives that were being written to or read from the respective LUN on the iSCSI SAN target are immediately adversely affected.

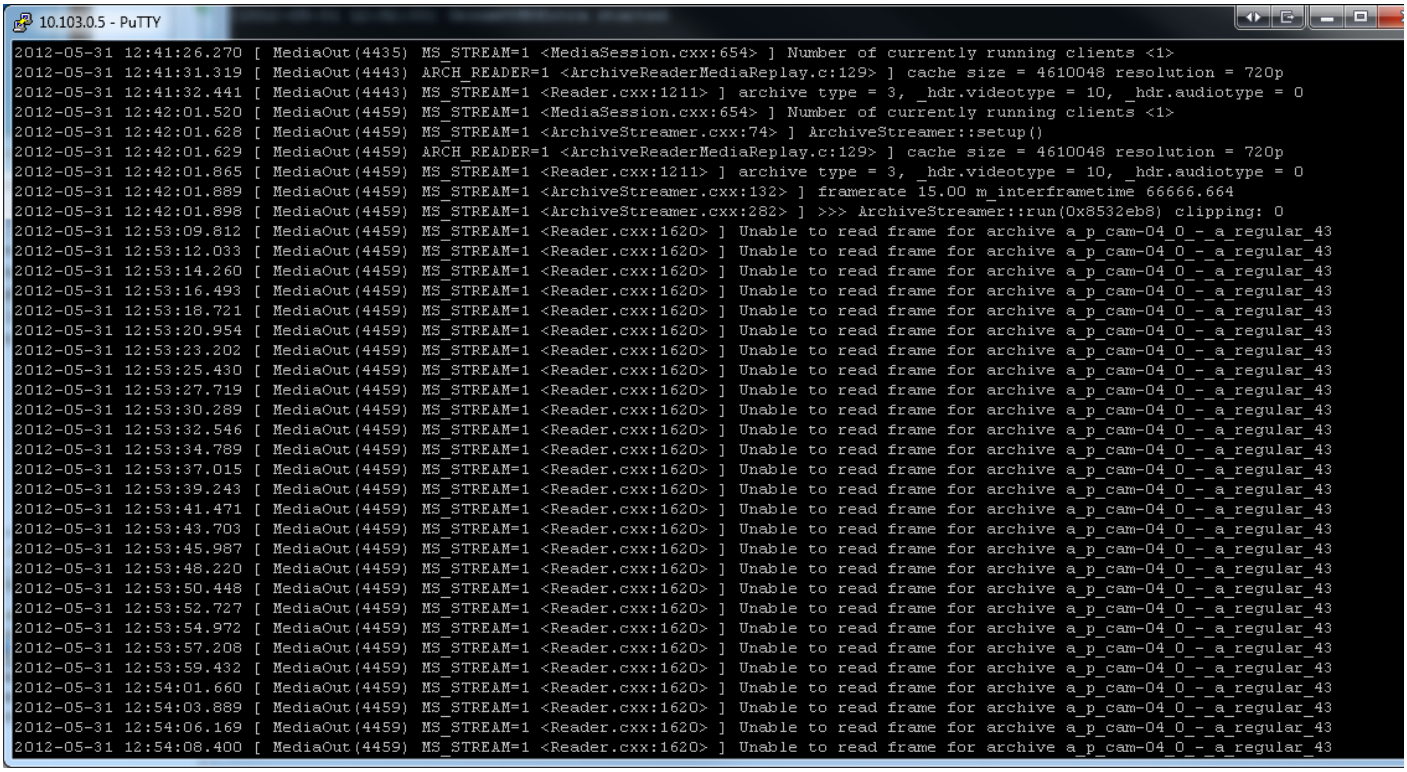
For archive data that is being written to the storage array, it is observed through the XCVRMAN log that the output queue buffers get full and begin to overflow because data in the queue is not being sent out from the queue. Figure 13 details this observation.

Figure 13 Scenario 3—Data Output



For archive data that is read from disk, the viewing client continues to display video content for as long as the input queue contains data, then the MEDIAOUT log, as well as the client display an error message indicating an error retrieving archive data. Figure 14 details this observation.

Figure 14 Scenario 3—Error Retrieving Archive Data



The hypervisor could either declare a state of Permanent Device Loss (PDL) or APD. The PDL condition is triggered when the respective LUN is not presented to the VM, either by unmounting the datastore and detaching the device (planned), or by unexpectedly deleting the LUN at the storage array. In both instances, the hypervisor can still communicate with the storage array and does receive valid sense codes. In the APD state, there is no communication with the storage array because the target is completely offline.

Note If the storage array is offline for durations greater than 30 minutes, it is highly likely that to recover, the iSCSI driver may need to be rescanned to rediscover the LUN targets. Additionally, the archives must be reselected from VSOM operator page to restart the respective archiver process and resume streaming video.

More Information

For more information about Cisco-related products, see the following resources:

- Internet Small Computer Systems Interface (iSCSI)

  http://tools.ietf.org/html/rfc3720
- Introduction to iSCSI White paper

  http://www.cisco.com/warp/public/cc/pd/rt/5420/prodlit/imdpm_wp.pdf
- tcp(7) Linux main page

  http://linux.die.net/man/7/tcp
- Permanent Device Loss and All-Paths-Down in vSphere 5.0

  http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2004684
- IP Video Surveillance Design Guide

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Video/IPVS/IPVS_DG/IPVSchap4.html
- TCP/IP Illustrated, Volume 1: The Protocols, 2nd Edition

  http://www.informit.com/store/product.aspx?isbn=0132808218
- Cisco SAFE Reference Guide: A Security Blueprint for Enterprise Networks

  http://www.cisco.com/en/US/docs/solutions/Enterprise/Security/SAFE_RG/SAFE_rg.html