iSCSI Implementation Guidelines
for Fault Tolerance and Load Balancing
using Temporary Redirection

Abstract

An approach for achieving fault tolerance and load balancing in iSCSI using the iSNS discovery mechanism or iSCSI discovery session and the temporary redirection mechanism is outlined here. This approach requires no change to iSCSI or other protocols that initiators and targets implement. But the manner in which initiators perform target discovery, support the temporary redirection mechanism, and recover from failed iSCSI sessions affects their ability to support this approach. This paper provides implementation guidelines for iSCSI initiators to follow to support this form of fault tolerance and load balancing.
1. Introduction

iSCSI can be used in a variety of configurations, including those in which the target is implemented as a distributed collection of nodes, or as a single node with multiple network interfaces. While fault tolerance and load balancing is not directly addressed by the protocol, the protocol does have two features that can be used to help build a fault tolerant solution: the target discovery process and the temporary redirection function. The iSCSI protocol’s login response process includes temporary redirection as a required feature. However, the recovery behavior after a redirected session fails or is explicitly terminated by an asynchronous event is not specified.

The paper details the behavior that iSCSI initiators may implement to enable a fault tolerant solution based on iSNS discovery or the discovery session and the temporary redirection features. This approach can be used to provide load balancing features as well. The initiator behavior outlined here is fully compatible and compliant with the iSCSI specification.

The paper first gives an overview of the approach, then details some examples of the sequence of events that occur when this approach is employed, and finally summarises the features initiators should implement to support temporary redirect based fault tolerant and load balancing systems.

2. Overview of the Approach

The solution relies on the initiator to establish a full feature phase session in a procedure that may take up to three steps, then to back up and repeat those steps in reverse order to recover the session if the underlying TCP connection fails or is terminated.

The initiator discovers the address or addresses of a target by either querying an iSNS server or by performing a discovery session to the "portal" configured for the target system. A portal, in iSCSI terminology, refers to an IP address and TCP port number pair. We term the address that an initiator connects to in order to perform a discovery session as the "discovery target portal." Both of these processes -- iSNS or discovery session -- return a set of portals for the target that we term the "initial target portals." Next, the initiator initiates a session to the target by trying to open a TCP connection to each of the initial target portals in sequence until one succeeds. Once connected, the initiator logs in. In its login response, the target may direct the initiator to a new portal via the iSCSI temporary redirection function. We term this new portal the "redirect portal" for the target. The initiator then closes the initial connection and attempts to initiate a session with the redirect portal. The negotiation of an iSCSI session to the redirect portal is the final step leading to a normal, full feature session between the initiator and the target, allowing data to flow.

If the connection fails, or is terminated by the target by an
asynchronous logout message, the initiator performs a sequence of actions to attempt to recover the session. Since this is session recovery, the initiator performs these recovery actions no matter what recovery level has been negotiated. The initiator first attempts to re-connect to one of the initial target portals learned from the iSNS query or during the discovery session, again trying each in sequence until one succeeds. If this fails, the initiator repeats the discovery phase, re-connecting to the discovery target portal and re-running the discovery session, or re-running the iSNS discovery procedure. After repeating the discovery process, the initiator follows the procedure it used at the time of initial connection, eventually connecting to a new redirect target for the target and recovering the full feature session.

Fault tolerance in this scheme is achieved by allowing the target system to direct the initiator to a specific network interface within a multi-homed iSCSI target system, or a specific network node in a clustered or distributed iSCSI target system for the duration of the session. In the event of failure, the target system can direct the initiator to a different, healthy interface or node, allowing the session to be recovered. The same mechanism provides load balancing by allowing the target to intelligently instruct the initiator to establish a new session with a less heavily loaded node.

This solution relies on initiator support of the iSCSI temporary redirect function. Additionally, the initiator must return to one of the initial target portals in the event that a connection fails or is terminated via an asynchronous logout message, and repeat the discovery process if that fails. The iSCSI specification does not dictate the precise recovery behavior for sessions established following a temporary redirection by a target. Some initiator implementations re-connect back to the same redirect portal after a connection to that portal fails. However, this behavior would prohibit the fault tolerant and load-balancing solution outlined here, and violates the spirit of the temporary redirection function. By returning to the redirect portal, the initiator is treating redirection as greater than temporary, but less than permanent.

3. Example Scenarios

To further illustrate how this mechanism works, this section presents the typical sequence of events that occur when a session is begun, when a target node or interface serving the session fails, and when a target decides to move an initiator to a different target node for load balancing.

Initial connection sequence of events:

1. Initiator performs the discovery procedure by using iSNS or executing a discovery session.

2. If iSNS is used, the initiator queries the iSNS server, which returns a set of portals for the target.
3. If the discovery session is used, the initiator opens a TCP connection to the discovery target portal, logs in and issues the "send targets" commands. The target responds with a list of target names and their associated portals. The initiator or user selects the portals associated with the specific target it is interested in establishing a session with. The initiator terminates the discovery session and closes the associated TCP connection.

4. Whichever discovery procedure is used, the initiator remembers the portals for this target as the "initial target portals".

5. The initiator iterates through the initial target portals list until it succeeds in opening a TCP connection to one of them.

6. The initiator then logs into the target, which may respond with a "target moved temporarily" redirect response, listing the redirect portal for the target. The initiator remembers this as the "redirect portal." The initiator then closes the TCP connection.

7. The initiator then opens a TCP connection to the redirect portal and logs in. The target accepts this login and the session proceeds to full feature phase.

8. Data flow begins.

Target node or interface failure sequence of events:

1. The initiator has an iSCSI session established and TCP connection open to the redirect portal. Full feature session in progress. Data is flowing.

2. The target fails.

3. The initiator detects the failure of the TCP connection with the target.

4. The initiator iterates through the list of initial target portals learned in the discovery process until it succeeds in opening a TCP connection to one of them.

5. If the initiator succeeds in connecting to one of the initial target portals, it executes steps 6 and 7 in the "Initial connection sequence of events" section.

6. If the initiator fails to connect to any of the initial target portals, it repeats steps 1 through 7 in the "Initial connection sequence of events" section.

7. Data flow resumes.

Overloaded target node sequence of events:
1. The initiator has a full feature iSCSI session established, and associated TCP connection open, to the redirect portal. Data is flowing.

2. The target terminates the session with an asynchronous logout message. The initiator closes the TCP connect.

3. If the asynchronous logout message PDU is type 1 (target requests logout), the initiator logs out, closes the TCP connection, and proceeds to step 4 in the "target node failure sequence of events" section.

4. If the asynchronous logout PDU is type 2 (target will drop connection), then Parameter2 (Time2Wait) specifies the time in seconds that the initiator should wait before attempting to re-login. The initiator should wait this time, then proceed to step 4 in the "target node failure sequence of events" section. The distinguished value of 0xFFFF may be used as an indication that the initiator should not re-login without the intervention of the administrator on the initiator. (The protocol provides no other way for the target to signal to the initiator that it does not wish it to re-connect.)

In these scenarios, the initiator is called on to determine when a TCP connection with a target has failed, and also when an attempt to open a new TCP connection to a target has failed. For both of these determinations, the iSCSI layer in the initiator can simply rely on the underlying TCP layer’s retransmission abort timeout mechanism, or it could implement timeouts of its own. The approach used, and the timeout values that the initiator selects, are highly implementation dependent. For example, some implementations allow applications to select the TCP abort timeout, while others do not. No matter what approach is taken, implementations may wish to make these two failure-determining timeout values configurable so that administrators may tune the system for operation in different environments.

4. Summary of guidelines for initiators

To summarize, the mechanisms that initiators should implement to support this approach for fault tolerance and load balancing are:

- Provide a target discovery mechanism by implementing either iSNS or the iSCSI target discovery session.

- Accept and act upon the iSCSI temporary redirect login response.

- If a session TCP connection to a redirect portal fails, try to re-connect to the initial target portals.

- If a session is terminated by the target with an asynchronous logout message, try to re-connect to the initial target portals.
5. Acknowledgements

Thanks to Bill Nowicki of Intransa, and Hari Mudaliar of Adaptec for their helpful comments on early revisions of this paper.

6. Security considerations

The initiator must perform authentication on every login.

7. References


[ISNS] Josh Tseng, Kevin Gibbons, Franco Travostino, Curt Du Laney, Joe Souza, \textquoteleft\textquoteleft Internet Storage Name Service (iNSS)\textquoteleft\textquoteright, \texttt{draft-ietf-ips-isns-17.txt}, Work in progress.

Authors Address

Robert E. Gilligan
Intransa, Inc.
2870 Zanker Road
San Jose, CA 95134
Email: gilligan@intransa.com
Phone: 408-678-8647

Rajkumar Velpuri
Intransa, Inc.
2870 Zanker Road
San Jose, CA 95134
Email: Rajkumar.velpuri@intransa.com
Phone: 408-678-8641

Lakshmi Ramasubramanian
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
Email: nramas@microsoft.com
Phone: 425-703-7559

Alan Warwick
Microsoft Corp.
One Microsoft Way
Redmond, WA 98052
Email: alanwar@microsoft.com
Phone: 425-706-0230

Matthew W. Baker
Intel Corporation
1501 South Mopac Expressway, #400
Austin, TX 78746
Email: matt.w.baker@intel.com
Phone: 512-732-1306