X3T10/95-314r0

From: Doug Hagerman
To: X3T10
Date: 8 September 1995
Subject: SCSI-3 High Availability Device Profile

Abstract: Proposal for a profile that describes requirements that apply to highly available systems based on the SCSI-3 interconnect.

1 PURPOSE

This profile describes the baseline SCSI requirements for building Highly Available storage subsystems based on SCSI bus hardware.

Both parallel and serial versions of SCSI-3 are covered. These requirements are addressed from the electrical, SCSI adapter/system, software and SCSI device perspectives. In particular this profile focuses on the requirements inherent in building SCSI multihost configurations. Multi-host SCSI is a requirement for building High Availability subsystems because it allows recovery from a host failure.

This profile is organized into two major sections. The first deals with general multi-host and high availability issues, ignoring any requirements that are specific to a given SCSI bus implementation. The second section deals with the specific requirements of parallel SCSI, SSA, and FC-AL implementations of the SCSI bus.

2 DESCRIPTION OF MULTI-HOST SCSI SYSTEMS

In order to support highly available storage subsystems, the use of more than one host initiator on the SCSI bus is a requirement. By allowing more than one host on the bus it becomes possible to move disk traffic from one host to another. To allow this "host failover" to occur, the requirements outlined below must be implemented in the SCSI systems, adapters and target devices.

Note that host failover is different from "controller failover", which describes the way that a controller (usually a RAID controller) failure is handled by a system. Controller

failover is usually implemented inside a controller cabinet, although it may extend out onto the SCSI bus.

There are a number of system issues that must be addressed to achieve a true high availbility system, including the failover of network traffic, shared access control, and management of the many issues related to system security, device naming, and access control. However, multi-host SCSI is a basic component that is needed because the system must have multiple paths to the same device on the bus.

Other configurations of storage interconnect topology may be used to achieve high availability. One method is to use dual ported devices with the ports dedicated to different hosts, while several network-based approaches are possible. However, multi-host access to the storage devices over a single bus is the configuration under discussion here.

The basic requirement is that the SCSI devices must be capable of coexisting on a SCSI bus with multiple hosts present.

This profile recognizes that the requirements for a host device are slightly different from those of a target device, and that diagnostic software has a slightly different set of requirements from the boot software or the run-time driver, but for the most part the requirements are common. Separate sections are reserved to describe any specific requirements for special cases.

3 DEFINITIONS

The following terms are defined for use in this profile.

3.1 High Availability SCSI Device

A SCSI device which meets the requirements of this profile.

3.2 High Availability SCSI Target

A SCSI device which meets the requirements of this profile, particularly in regards to those requirements are specifically applicable to SCSI targets.

3.3 High Availability SCSI Initiator

A SCSI device which meets the requirements of this profile, particularly in regards to those requirements are specifically applicable to SCSI initiators.

4 GENERIC SCSI-3 MULTI-HOST REQUIREMENTS

This section describes the generic requirements for multi-host high availability SCSI systems. These requirements apply equally to Fast-20, SSA, and FC-AL implementations of SCSI-3.

4.1 Generic SCSI-3 Physical Requirements

The basic physical requirement is that the SCSI bus should be capable of supporting multiple hosts. This means that it should be possible to construct a bus configuration that includes multiple hosts, devices, subsystems or other components without the need to perform "objectionable" levels of hardware manipulation. An example of an objectionable manipulation would be if a host required that the enclosure be opened to remove bus terminators before it could be used in the multi-host configuration.

In most installations the SCSI bus for a multi-host configuration should be expected to be at least several meters in length. (Clearly special enclosures with multiple hosts would require shorter minimum lengths.) In those situations where it is desired to maintain the hosts on separate power grids, the SCSI bus might need to be several kilometers in length.

The installation should protect against SCSI bus configurations that are possible to construct with available hardware but that do not operate satisfactorily because of electrical or other reasons. It is desireable that the system have a means of performing an automatic verification of the configuration. This means that one should not construct a Fast-20 single ended system with a 20 meter bus length, and that the system should detect it if it happens.

The SCSI bus should support the "plugging" (removal or insertion of a device on the SCSI bus) of devices or hosts without disturbing the other devices on the bus. This means that the cabling should regain full continuity after the plugging operation is complete. This may require the use of "Y" cables for Fast-20 SCSI-3 or "Loop Resiliancy Circuits"

for FC-AL or "blank device enclosures" for.

4.2 Hot Vs Warm Vs Cold Plug

Depending on the level of plugging support desired, the SCSI bus should maintain continuity for all or most of the period during which devices are plugged. The levels are defined and discussed below.

There is no standard terminology to describe device plugging. "To plug" means either to remove or to insert a device, including the removal or connection of any mechanical interlocks or retainers, and the disconnection or connection of the electrical bus interface to the device. In 1990 IBM proposed the term "live plugging" for the case that is now commonly called "hot plugging". Other possibilities identified at that time are listed here for reference.

0 No power applied anywhere during the device plugging operation.
1 Concurrent maintenance: device power off, others on, but bus idle.
2 Hot plugging: device power on, others on, bus idle.
3 Live plugging: device power on, bus in use by other devices.
4 Yanking: bus in use by device being plugged.

In the case of highly available systems, either case 3 or case 4 makes sense. It is undesireable to require that the bus be idle during the plug operation, so cases 0, 1, and 2 are not considered further in this profile.

For the purposes of this profile, "hot plugging" refers to case 3, above.

4.3 Generic SCSI-3 Electrical Requirements

The SCSI bus and associated electronics should support hot plugging. There may be a need for longer pins in some connectors in order to properly sequence the connection and disconnection of the power, ground, and data circuits.

The driver and receiver electronics should be able to withstand the hot plugging operation. The signal lines should be well-behaved during the hot plugging operation. Other devices and in-progress transfers should not be disturbed by

the hot plugging operation.

The driver and receiver electronics should be well-behaved during power cycles. No glitches or other irregular signals may be caused on the bus as a result of the application or removal of power to the device during or in preparation for the plug operation.

After power is initially applied to a device after it is plugged, the unit attention flag should be set. The SCSI unit attention flag should be maintained on a per-initiator basis in the device.

4.4 Generic SCSI-3 Logical Requirements

The device ID mechanism used by the bus should be fully supported by the multi-host environment. If the host software requires fixed bus IDs, then the devices on the bus should implement fixed bus IDs. If the host software supports dynamic device addressing, the devices on the bus may implement dynamic addressing.

In order to be able to uniquely identify a device regardless of where it is in a configuration, or after swapping, or in the case of multiple access paths, all devices should support the vital product data unit serial number page.

Since general bus resets can be extremely costly in terms of performance across the entire system, they should be issued only as a last resort. Only hosts should issue general resets.

Since specific device resets (BUS DEVICE RESET) may interfere with device activity that was started by another host, device resets should be sent only by hosts.

All devices should handle all possible incoming messages at all times. In particular, console microcode used during the host initialization process should implement the complete message protocol because other traffic will be active on the bus during host initialization.

All initiators should renegotiate any bus options (e.g. wide SCSI) with any device that may have been replaced or power cycled since it was last used. This should not be done on every command, but after a host determines that a bus event has occurred, as detected by the use of the Unit Attention flag.

All initiators should renegotiate any bus options such as wide or synchronous before issuing the INQUIRY command.

4.5 Generic SCSI-3 Command Requirements

Devices should implement all the mandatory SCSI commands for the device type they report. Optional commands that are not implemented should be handled properly according to the SCSI standard.

4.6 Generic SCSI-3 Target Device Requirements

Devices should properly handle bus resets that may occur at any time. Setup information should not be carried across a bus reset (except for saved mode parameters as described in SCSI) because a newly added host cannot predict the earlier information.

Devices should maintain mode pages on a per-LUN basis. This is needed because hosts view each LUN as a separate device.

Devices should properly support simultaneous hosts. Devices should be able to accept and process commands from multiple initiators at any bus IDs without hanging the bus, violating the SCSI standard, or crashing or hanging themselves.

Devices should properly handle device reservation (using either RESERVE or PERSISTENT RESERVE) in a multi-host environment.

Devices should support tagged commands in a timely manner.

4.7 Generic SCSI-3 Initiator Device Requirements

Hosts should minimize the number of bus reset operations that they initiate. This means that a host should attempt to never reset the bus during initialization, normal processing, and shutdown. A bus reset should be sent only when it is determined that no other method of restarting the bus is possible. Prior to resetting the bus the host should coordinate with other hosts on the bus.

4.8 Generic SCSI-3 System Level Requirements

Device mode pages should be coordinated between all hosts and devices on the bus. Devices are not required to maintain mode pages on a per-initiator basis, so all hosts should be able to operate with the same mode page setup on each device. Each device may have different mode parameter values, but the values for a given device apply across all hosts.

Device resource locking should be controlled using RESERVE or PERSISTENT RESERVE commands. The locking should be coordinated between all hosts and devices on the bus.

SCSI Device reservations should be coordinated between all the hosts in the system. Since the status of a reservation may change upon removal of device power, the hosts should coordinate the reservations between themselves. The persistent reservation option may be used to improve this coordination.

Bus ID assignments should be coordinated between all hosts in the system. This applies both in the case of fixed bus IDs and dynamic IDs.

The use of BUS DEVICE RESETs should be coordinated between all the hosts in the system.

5 BUS-SPECIFIC REQUIREMENTS

The following sections define the high availability requirements for each of the three implementations of SCSI-3.

5.1 Parallel SCSI-3 Requirements

The SCSI system, including all devices, components, hardware and software, shall conform with all requirements of the latest revision of the SCSI-3 Parallel Interface standard (SPI). Revision 15a is current as of the date of this profile.

Specific additional exclusions and expansions to the SPI standard are described in this profile.

5.1.1 Parallel SCSI-3 Physical Requirements -

In order to allow devices to be removed from the SCSI bus without interrupting bus activity, the cable plant shall provide electrical continuity and bus termination when a device is removed from the bus.

SCSI devices and adapters shall conform to one of the following two cabling options. Refer to Note 3, SPI (page 8).

The two options may be mixed on a single SCSI bus as long as the continuity requirement is met.

1. The preferred option is the one connector option. The enclosure shall be implemented with a single external SCSI connector. In order to remove such a device from the SCSI bus without interrupting bus activity, the cable plant shall be equipped with "Y" SCSI cables. SCSI bus terminators shall not be installed inside the enclosure, but shall be connected directly to the SCSI bus itself. The stub length of the connector and cable inside the enclosure shall meet the requirements of SPI sections 6.4 and 6.5.

2. A less desireable option is the two connector option. The enclosure shall be implemented with two external SCSI connectors. In this case the SCSI bus enters and exits the device using the two connectors, and the internal wiring is arranged to minimize the stub length caused by the device connection. Such an

enclosure shall meet the requirements of SPI section 5.2.

It should be noted that enclosures that use the two connector option cannot be removed from the SCSI bus without disrupting bus activity. However, an enclosure suitable for high availability systems may be wired this way if it allows the devices it contains to be removed without disruption of the SCSI bus, and if it provides internal bus continuity when devices are removed from the enclosure.

The SCSI cable plant shall meet the recommendations of SPI Annexes D and F. Cable lengths for a complete system shall not exceed the following values.

Single Ended:	
Up to 5 Megatransfers per second	6 meters
5 to 10 Megatransfers per second	3 meters
10 to 20 Megatransfers per second	1.5 meters
Differential:	
All speeds	25 meters

Since the short cable lengths associated with higher clock speeds may be impractical for high availability systems, it is expected that most such systems will use the differential signalling alternative.

5.1.2 Parallel SCSI-3 Electrical Requirements -

A device shall not be terminated internally or terminated in such a way that precludes it from occupying any position on the SCSI bus.

Switchable terminators may be used if there is a mechanism for them to be disabled, such as by software or a jumper.

Since bus extenders and converters (eg single-ended to differential) terminate an electrical bus segment, they shall provide bus termination. This termination shall be external to the extender or converter.

Terminator power shall be supplied as described in SPI section 7.3, except that "optional internal terminators" shall not be used.

Signals on the SCSI bus shall not be disrupted during power transitions. SCSI targets and initiators shall retain their high impedance state during power cycles. Refer to SPI

sections 7.1.2 and 7.2.2.

A high availability system requires that devices be inserted or removed from the bus during bus activity to other devices. This is needed to allow system maintenance, repair, and reconfiguration while normal operation continues. In order to insure glitch free insertion and removal of devices onto to, or off of the SCSI bus, SCSI devices shall conform to SPI paragraph A.4, "Current I/O Process Allowed During Insertion or Removal".

The system software shall arrange to prevent bus activity to the device that is to be plugged (the device becomes inactive on the SCSI bus), and the system hardware shall guarantee that the device power and ground connections are made before the SCSI signal lines, in conformance with the requirements of SPI paragraph A.4.

To meet the requirements of SPI paragraph A.4, one of the following two schemes shall be implemented:

1. The required signal sequencing may be done by using staggered pins in the connector.

2. The enclosure hardware or software may cooperate with the host software to achieve bus activity to the device. The required signal sequencing may be done by using electrical or mechanical switching devices that isolate the device being plugged from the SCSI bus.

5.1.3 Parallel SCSI-3 Message And Command Requirements -

High availability SCSI devices, including targets and initiators, shall return a status of CHECK CONDITION with sense key of ILLEGAL REQUEST for any unsupported command.

High availability SCSI devices, including targets and initiators, shall return a status of CHECK CONDITION with sense key of ABORTED COMMAND/MESSAGE ERROR for any unsupported message.

Upon host bootup, the console code of high availability SCSI initiators shall negotiate synchronous speed and wide data transfer setup before attempting to execute any SCSI command that enters a data phase (including the INQUIRY command) to any target. Initiators shall assume that any device on the bus may have been powered on, reset, or have changed to a fast

or wide mode since the last time it was used by the initiator.

This DOES NOT mean that these negotiations should be done before each command, rather the console (NOT runtime drivers) should do this before the first I/O during the boot or shutdown sequence.

High availability SCSI devices shall support the vital product data unit serial number page.

5.1.4 Parallel SCSI-3 Target Device Requirements -

High availability SCSI target devices shall not issue SCSI bus resets.

Targets in Multi-host systems shall be able to accept and process commands from multiple initiators.

Targets shall accept and process commands from initiators located at any bus ID.

When a target that is holding data in a cache before writing it to non-volatile storage receives a bus reset, it shall write the cache contents to the media before processing the reset.

Targets shall maintain the following on a per-initiator basis.

Synchronous negotiated state.
Width negotiated state.
Contingent Allegiance state.
Unit attention flag.

The Unit Attention Condition shall indicate whether if the mode parameters in effect for this initiator have been changed by another initiator, or if the mode parameters in effect for the initiator have been restored from non-volatile memory, or if any of the normal SCSI-2 Unit Attention Condition conditions apply.

If a target device supports LUNs, then mode pages shall be maintained on a per-LUN basis.

Targets shall support the RESERVE and RELEASE SCSI commands. These commands allows hosts to allocate devices with exclusive access. [question of PERSISTENT RESERVE...]

Targets shall support the following mechanisms to clear device reservations:

RELEASE Command
BUS DEVICE RESET Message
SCSI Bus Reset
Power Down/Remove

Targets that are reserved by an initiator shall accept and process the following commands received from any initiator. All other commands shall be failed with a SCSI status of RESERVATION CONFLICT.

INQUIRY
REQUEST SENSE
PREVENT ALLOW MEDIA REMOVAL (Bit set 0) (removable devices)
RELEASE

If the RELEASE command is received from the initiator that has the outstanding reservation, the reservation is cancelled. If the RELEASE command is received from another initiator, the command is failed with a SCSI status of RESERVATION CONFLICT.

High Availability SCSI target devices shall support Tagged Command Queuing.

High Availability SCSI target devices shall support drive based Bad Block Replacement (BBR) as described in SCSI-3.

High Availability SCSI target devices shall implement a reselection retry algorithm that limits the amount of bus time spent attempting to reselect a non-responsive initiator.

5.1.5 Parallel SCSI-3 Initiator Device Requirements -

It is particularly important that host and host adapter designs comply with the requirements for external bus termination. It is also important to consider the system implications of the choice between the single-connector and dual-connector options.

If these requirements are not met, the SCSI system is limited to two hosts that cannot be hot-plugged. This is not adequate for a high availability system.

Because bus activity is expected to continue during the power sequencing, removal, replacement, and reboot procedure on a failed host in a high availability system, there is no distinction between the requirements placed on the console

micr@ Èe, the host adapter microcode, and the normal runtime driver software environment. Every device on a high availability SCSI bus shall meet all the requirements at all times. This is a notable difference from a single-user system where boot-time discrepancies from normal SCSI usage are common.

If a host is halted by an operator command (such as a console halt command) the SCSI bus host adapter shall not stall in a state that prevents the other devices on the SCSI bus from continuing normal operation.

The host shall support the processor type device mandatory requirements per ANSI SCSI-3.

High availability SCSI initiators, including system consoles, and adapter microcode, and mainline driver code, shall not issue SCSI bus resets except when the SCSI bus is "hung". The definition of "hung" is system dependent, but the following procedure is recommended.

The initiator that suspects that a target is hung should first attempt to issue an INQUIRY command to the target. If the command goes through the required bus phases then the bus itself is assumed not to be hung. If this fails, the initiator should attempt to coax the target to MESSAGE OUT phase [describe how ???], and then send an ABORT message. If that fails then the initiator should attempt to send the BUS DEVICE RESET MESSAGE, if that fails the initiator should communicate with the other cooperating hosts on the bus to determine whether it is ok to issue a bus reset.

The SCSI bus reset signal is extremely disruptive to in-progress bus activity and can be require lengthy recovery activity, particularly in the case of systems with tape drives and highly cached storage subsystems.

It may be better to remove a suspect device from the bus than to attempt on-line recovery. The reset signal shall be used only as a last resort.

A third party reset is a reset that an initiator detects that was generated by another device. The initiator shall be able to recover from a single or repeated third party SCSI bus resets. The initiator shall not take longer than 60 seconds to recover from a single SCSI bus reset or from the last of any series of resets.

(Note: The 60 second requirement is intended to define a guideline for the amount of time the SCSI I/O subsystem may take to recover from a bus reset. The actual recovery time

for the entire system may be longer than 60 seconds, depending on a number of factors including the number of spindles on the bus, whether failover actions occur and whether or not the file system recovery is fast or slow.)

Event during the period when a system is starting up or shutting down in a Multi-host environment, it is still possible for other initiators to select the system as a target. This selection shall be treated as a normal event and handled in such a way that will allow the currently executing boot or shutdown activity to complete without error.

Three optional solutions to this situation may be used.

1. Disable selection as a target in the console or adapter card. This option is most appropriate for adapters that have little intelligence on them.

2. Enable selection as a target and support the INQUIRY, TEST UNIT READY and REQUEST SENSE SCSI commands. Use of this option implies that until the host software has completed its boot process, the console microcode shall be able to respond to and process these commands, and shall either completely implement all of the SCSI message protocol or correctly REJECT any unsupported SCSI bus messages received.

3. Option 3: Enable SELECTIONs and return SCSI status of BUSY, and then return the COMMAND COMPLETE message.

Initiators shall not use all the tag queue depth in a device. The initiator shall reserve some number of tag queue elemenets so other initiators may still send commands (such as INQUIRY) to the device while it is in use by another initiator.

5.1.6 Parallel SCSI-3 System Level Requirements -

It is recommended that the use of "Y" cables and the single-connector option be chosen for all designs. This approach maximizes configuration flexibility and provides the opportunity to maximize system availability.

It is recommended that the mass storage devices in a high availability system be housed in a self-contained enclosure that is separate from the host enclosures. By providing independent power supplies and cabinet services a storage subsystem can provide SCSI block data service to more than one

host system with minimal disruption if one host system malfunctions.

It is recommended that dual redundant independent power supplies be provded in the storage enclosure since the maintenance of electrical power to the devices is critical in maintaining high availability. If data redundancy is distributed between more than one storage subsystem this is not as important.

If storage devices are housed within a host system enclosure, it is critical to insure that the external SCSI bus connectors to these internal devices do not violate the SCSI bus stub length requirements when connected to a "Y" cable.

All high availability SCSI devices shall implement at least the "SCAM tolerant" level of SCAM as described in SPI Annex B.

In order to have more than one initiator on the SCSI bus there shall be a cooperative method of handling the SCSI ID assignments on all the devices on the SCSI bus. Either of the following may be used.

1. All devices on the bus have SCSI IDs assigned in advance by the system administrator and set by switches or jumpers. Each initiator shall have a unique SCSI bus ID.

2. The various levels of SCAM support shall be implemented according to SPI Annex B.

Console, host adapter, and device diagnostics, self tests, and boot sequences shall run correctly on an active SCSI bus. Console, host adapter, and device firmware shall not effect active I/O on the bus.

High availability multi-host systems shall have a mechanism to coordinate the access to shared data. SCSI provides such a mechanism in the RESERVE and RELEASE commands, but the details of data access coordination is vendor specific.

High availability multi-host systems shall have a mechanism to coordinate the mode page settings on shared devices. The details of mode page coordination is vendor specific.

Particular examples of mode page values that shall be maintained on a system-wide basis include:

Default block size
Read/Write Error recovery page

Cache control page
Disconnect/Reconnect page

5.2 FC-AL SCSI-3 Requirements

[compare to list of proposed low cost FC-AL implementation]

5.2.1 FC-AL SCSI-3 Physical Requirements -

FC-AL requires active participation by each loop member to maintain loop continuity. To avoid a single point of failure, FC-AL systems use a pair of independent redundant loops.

5.2.2 FC-AL SCSI-3 Electrical Requirements -

To minimize the disruption of the loop during device plugging, the subsystem shall use a multiplexor of some type to bypass a device during the plugging operation. This may be located at the device or in a physically remote location that is topologically near the device. The latter choice allows all the active redundancy hardware to be localized to a single physical location which may make it easier to provide redundant electrical power to the critical components.

5.2.3 FC-AL SCSI-3 Message And Command Requirements -

The dynamic addressing scheme used by FC-AL shall be supported by the devices and hosts used in a redundant system. One option is to have the addresses set by hardware (switches or jumpers); another is to have full cooperative support for dynamic device addressing in the host systesm.

5.2.4 FC-AL SCSI-3 Target Device Requirements -

An FC-AL device shall meet the requirements of dual port SCSI.

5.2.5 FC-AL SCSI-3 Initiator Device Requirements -

TBD

5.2.6 FC-AL SCSI-3 System Level Requirements -

The FC-AL bus redundancy is based on redundant loops, The host shall support traffic in both directions from either port.

5.3 SSA SCSI-3 Requirements

5.3.1 SSA SCSI-3 Physical Requirements -

SSA requires active participation by each loop member to maintain loop continuity. To avoid a single point of failure, SSA systems use a pair of interdependent redundant loops.

5.3.2 SSA SCSI-3 Electrical Requirements -

The SSA architecture results in the ability of SSA devices to be removed from a loop without disrupting the electrical status of the loop.

The removal of more than one device from an SSA loop (or a partially-populated cabinet) may cause the isolation of some of the devices. For a high-availability system, the isolation shall be prevented by the use of dummy plug-in modules or a multiplexor of some type to bypass a missing device.

5.3.3 SSA SCSI-3 Message And Command Requirements -

The dynamic addressing scheme used by SSA shall be supported by the devices and hosts used in a highly available system.

5.3.4 SSA SCSI-3 Target Device Requirements -

SSA devices shall meet the requirements of dual port SCSI. In the case of a failure of a device in the loop that causes traffic coming to a device via one of its ports to be interrupted, the device shall support the failover mechanism that causes incoming traffic to a device to be reflected back using the outgoing connection on that port.

5.3.5 SSA SCSI-3 Initiator Device Requirements -

SSA hosts shall meet the requirements of dual port SCSI. In the case of a failure of a device in the loop that causes traffic via one path to be interrupted, the host shall support the failover of that traffic to its other port.

5.3.6 SSA SCSI-3 System Level Requirements -

Since SSA bus redundancy is based on dual-ported hosts that fail over to the alternate path, the SSA configuration shall be a loop. The other SSA topologies shall not be used in high availability systems.