Working Draft

X3T10/TBDTR

Revision 3
December 31, 1996

Information Technology - Profile for Parallel SCSI Components Used in High Availability Environments

Draft proposed American National Standard

This is a draft proposed Technical Report of Accredited Standards Committee X3. As such this is not a completed Technical Report. The X3T10 Technical Committee may modify this document as a result of comments received during public review and its approval as a technical report. Use of the information contained here is at your own risk.

ABSTRACT
This Technical Report defines a profile for the use of parallel SCSI equipment in environments where high system availability is required. A recommended solution is included.

Technical Editor:
Douglas Hagerman
Digital Equipment Corporation
SHR3-2/C5
334 South Street
Shrewsbury MA 01545
Voice: 508-841-2145
FAX: 508-841-6100
EMail hagerman@starch.enet.dec.com

Reference number
ISO/IEC ***** : 199x
ANSI X3.*** - 199x
Printed 01/10/97

Other Points of Contact:

	X3T10 Chairman	X3T10 Vice-Chair
	John Lohmeyer	Lawrence Lamers
	Symbios Logic Inc	Adaptec
	4420 ArrowsWest Drive	MS 293
	Colorado Springs, CO 80907-3444	691 South Milpitas Blvd
		Milpitas CA, 95035
Voice:	719-533-7560	408-957-7817
Fax:	719-533-7036	408-957-7193
Email:	john.lohmeyer@symbios.com	ljlamers@aol.com

X3 Secretariat
Lynn Barra
Administrator Standards Processing
X3 Secretariat
1250 Eye Street, NW Suite 200
Washington, DC 20005

Voice: 202-626-5738
FAX: 202-638-4922
Email: lbarra@itic.nw.dc.us

Reflector

Internet address for subscription to the X3T10 reflector:	majordomo@symbios.com
Note should contain a line stating...	subscribe scsi *your email address*
Internet address for distribution via X3T10 reflector:	scsi@symbios.com
X3T10 Bulletin Board	719-533-7950
FTP Site:	ftp.symbios.com /pub/standards/io/x3t10
Web sites:	http://www.x3.org/x3t10 or http://www.symbios.com/x3t10 http://www.ssaia.org

Document Distribution
Global Engineering
15 Inverness Way East
Englewood, CO 80112-5704

Voice: 303-792-2181
or: 800-854-7179
FAX: 303-792-2192

PATENT STATEMENT

CAUTION: The developers of this Technical Report have requested that holders of patents that may be required for the implementation of the Technical Report disclose such patents to the publisher. However, neither the developers nor the publisher have undertaken a patent search in order to identify which, if any, patents may apply to this Technical Report.

As of the date of publication of this Technical Report and following calls for the identification of patents that may be required for the implementation of this Technical Report, no such claims have been made.

No further patent search is conducted by the developer or the publisher in respect to any Technical Report it processes. No representation is made or implied that licenses are not required to avoid infringement in the use of this Technical Report.

Contents

Figures

X3's Technical Report Series

This Technical Report is one in a series produced by the American National Standards Committee, X3, Information Technology. The secretariat for X3 is held by the Computer and Business Equipment Manufacturers Association (CBEMA), 1250 Eye Street, NW Suite 200, Washington DC 20005.

As a by-product of the standards development process and the resources of knowledge devoted to it, X3 from time to time produces Technical Reports. Such Technical Reports are not standards, nor are they intended to be used as such.

X3 Technical Reports are produced in some cases to disseminate the technical and logical concepts reflected in standards already published or under development. In other cases, they derive from studies in areas where it is found premature to develop a standard due to a still changing technology, or inappropriate to develop a rigorous standard due to the existence of a number of viable options, the choice of which depends on the user's particular requirements. These Technical Reports, thus, provide guidelines, the use of which can result in greater consistency and coherence of information processing systems.

Foreword

The traditional use of SCSI is for the connection of storage devices to small desktop systems or workstations. Over a period of several years SCSI has matured to the point where it is the preferred method of interconnecting almost all high-performance disk drives as well as a large fraction of related devices such as CD-ROM and tape drives. In addition the SCSI interface has been improved in both performance and reliability.

As a result of these developments SCSI is now seen as an interconnect that is suitable for use on high-end computer systems where a high degree of system availability is required. The use of SCSI components in such demanding systems puts additional requirements on how the SCSI components operate, primarily in the form of "good behavior" requirements that allow the system to continue to operate even in the event of certain failures.

Because of the developmental history of SCSI, many devices already in the field do not meet this higher level of requirements, and even new devices and components may not meet them unless there is a specific description of what the requirements are and why they are important. The purpose of this Technical Report is to provide guidance to SCSI device implementers so that they may make implementation choices that maximize the usefulness of the devices in High Availability systems.

This Technical Report includes an overall description of how SCSI may be used to create High Availability systems, lists the generic goals and top-level requirements on components to be used in such systems, and provides a detailed set of implementation requirements that reflect the generic goals.

This Technical Report was developed by Task Group X3T10 of Accredited Standards Committee X3 during 1996-97.

Requests for interpretation, suggestions for improvement and addenda, or defect reports are welcome. They should be sent to the X3 Secretariat, Information Technology Industry Council, 1250 Eye Street, NW, Suite 200, Washington, DC 20005-3922.

This Technical Report was processed and approved for submittal to ANSI by Accredited Standards Committee on Information Processing Systems, X3. Committee approval of the Technical Report does not necessarily imply that all committee members voted for approval. At the time it approved this Technical Report, the X3 Committee had the following members:

James D. Converse, Chair

Donald C. Loughry, Vice-Chair

Joanne M. Flanagan, Secretary

Technical Committee X3T10 on I/O Interfaces, which reviewed this Technical Report, had the following members:

John B. Lohmeyer, Chair Lawrence J. Lamers, Vice-Chair Ralph Weber, Secretary

[names to be added]

Introduction

This Technical Report is divided into the following clauses and annexes.

Clause 1 defines the scope of the document.

Clause 2 specifies the normative references including relevant SCSI-3 standards, and explains the relationship between Technical Reports and Standards.

Clause 3 defines the definitions, symbols and abbreviations.

Clause 4 contains the conventions.

Clause 5 defines High Availability SCSI Systems.

Clause 6 defines Multi-Host SCSI Systems.

Clause 7 defines the Fundamental Requirements for High Availability Multi-Host SCSI Systems.

Clause 8 defines the System Level Requirements.

Clause 9 defines the Physical Requirements, including enclosure and connector requirements.

Clause 10 defines the Electrical Requirements, including line driver and receiver requirements and terminator requirements.

Clause 11 defines the Logical and Command Requirements, including the requirements that apply to the various SCSI-3 data phases.

Clause 12 defines the Target Device Requirements.

Clause 13 defines the Initiator Device Requirements.

Clause 14 defines the Requirements that apply to specific Device Types.

Clause 15 defines the effect on existing standards.

X3 Technical Report for Information Technology -

Profile for SCSI Components Used in High Availability Environments

1. Scope

This Technical Report defines a profile for the use of parallel SCSI components in systems where high system availability is obtained by the use of multiple hosts and multiple devices connected by a single SCSI bus. A profile of SCSI-3 features needed to construct such systems is included.

2. References

X3's Technical Reports are not Standards and thus do not contain binding requirements. Therefore they do not contain normative references. Any statements herein that appear to contain requirements (especially instances of sentences with the verb "shall") are advisory in nature. Not following these advisory statements will likely result in implementations that do not accomplish the stated goals of this Technical Report. Thus, the verbs "may", "should", and "shall" are to be interpreted as follows:

- "may" means "is allowed to",
- "should" means "is recommended to", and
- "shall" means "is recommended to in order to accomplish the goals of this Technical Report".

This Technical Report is written with the understanding that it is for use with SCSI components as described in the set of SCSI-3 Standards.The SCSI architecture, as well as references to all the relevant standards, is described in the SCSI-3 Architecture Model (SAM) (X3T10-994D). SAM defines the functional partitions and specifies a model for SCSI-3 I/O system and device behavior which applies to all SCSI interconnects, protocols, acces methods and devices. The parallel SCSI bus is described in a pair of standards. These are the SCSI-3 Parallel Interface (SPI) (X3T10-885D) and the SCSI-3 Interlocked Protocol (SIP) (X3T10-856D). Other features of SCSI-3 are described in other related standards.

2. Definitions, Symbols and Abbreviations

2.1 Definitions

[claim by Gene Milligan that these definitions don't add anything, and leave out relationship to component failure]

2.1.1 High Availability SCSI System

A computer system that uses SCSI-3 devices and components to meet the requirements of this Technical Report. The purpose of a High Availability SCSI System is to maximize the availability of user data and to maximize the availability of the associated processing systems.

2.1.2 High Availability SCSI Device

An SCSI device which meets the requirements of this Technical Report. The purpose of a High Availability SCSI device is to operate within a subset of the SCSI-3 requirements so as to maximize the performance and availability of the complete system. A SCSI device may be either a target or an initiator.

2.1.3 High Availability SCSI Target

An SCSI device which meets the requirements of this Technical Report, particularly in regards to those requirements that are specifically applicable to SCSI targets.

2.1.4 High Availability SCSI Initiator

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

2.1.5 Failover

The process of recovering from the failure of a component by transferring its workload to another component.

2.1.6 Host Failover

An event that occurs in the case of the failure of one of a set of redundant host computers. The failure of the host causes its workload to be transferred to one or more of the other hosts.

2.1.7 Storage Device Failover

An event that occurs in the case of the failure of one of a set of redundant storage devices. The failure of the device causes its I/O load to be transferred to one or more of the other storage devices.

2.1.8 Controller Failover

An event that occurs in the case of the failure of one of a set of redundant storage subsystem controllers. The failure of the controller causes its I/O load to be transferred to one or more of the other controllers.

2.1.9 Standby Mode

An operation mode of a host, controller, or storage device in which the unit is available for use but does not participate in supporting the normal workload of the system. In this case the unit is not in contention for system resources.

2.1.10 Active Standby Mode

An operation mode of a host, controller, or storage device in which the unit is used during normal operation but is also available for use as a spare in case of the failure of another component. In this case the unit is in contention for system resources during normal operation, and the system needs a method for preventing potential damage to user data that could occur if two components attempt to service the same user request at the same time.

2.1.11 Device Plugging

The addition or removal of a device from a storage subsystem. In some cases this may be done with the system active, while in other cases the system must be idle or powered off before the device can be plugged.

2.1.12 Y Cable

A SCSI cable that provides a stub connection in mid-cable.

2.1.13 Console

The mechanism used by human operators to communicate with a system component for control purposes including startup, configuration, and other operations that require low-level access to the component.

Abbreviations

GByte gigabyte
MByte megabyte

3. Conventions

Certain words and terms used in this Technical Report have a specific meaning beyond the normal English meaning. These words and terms are defined either in the glossary or in the text where they first appear. Lower case is used for words having the normal English meaning.

All numbers used in this Technical Report represent decimal values. Numbers having a fractional part are indicated with a comma (e.g., two and one half is represented as "2,5”). Numbers having a value exceeding 999 are represented with a space (e.g., twenty four thousand two hundred fifty five is represented as "24 255").

4. High Availability Systems

This Technical Report describes the baseline SCSI requirements for building High Availability computer systems based on parallel SCSI bus hardware. These requirements are addressed from the electrical, SCSI bus host adapter, software and SCSI device perspectives. The purpose of such a system is to maximize the availability of the users' data in the presence of various potential failures. The basic feature of a High Availability system is that most or all single-component failures may be survived because redundant components are built into the configuration. In particular, a High Availability system in general has two or more copies of the users' data stored on non-volatile media, and has two or more host processors any of which can execute user application programs. This is not meant to imply that data mirroring is the only way to implement a High Availability system; other methods of providing data redundancy may also be used.

Depending on the sophistication of such a system the multiple copies of user data and the multiple host processors may be used in various modes. If the system supports a "standby" mode of operation, the redundant data or processors are only used in the case of a failure. If the systems supports an "active standby" mode of operation, the redundant data or processors are used during normal operation, and in the case of a failure operation continues with degraded performance.

5. Description of Multi-Host SCSI Systems

There are several approaches to the construction of High Availability computer systems. This Technical Report describes an approach that uses multiple hosts and multiple devices connected by a single parallel SCSI bus. Systems of this type are called "Multi-Host High Availability" configurations in this Technical Report.

(Note that systems of this type may have more than one SCSI bus, either by using separate busses for the host-to-subsystem connection and the subsystem-to-device connection or by using bus extenders which use multiple separate physical busses in a configuration that appears to be a single logical SCSI bus.)

By using more than one host on the bus it becomes possible to move I/O traffic from one host to another. By using more than one storage device on the bus it becomes possible to maintain multiple copies of data. These two features allow "host failover" to occur in case a host fails, and "storage device failover" to occur in case a storage device fails. To allow these failover operations to occur, the requirements described in this Technical Report 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 because to obtain high failover performance it is usually necessary that there be a direct high-speed connection between the controllers.

A requirement is that a High Availability system must support the removal and insertion of devices from the SCSI bus while the system is in operation. This may be needed for normal service or in the case of a failure. In either case the system must continue to operate, even if only at a degraded level of performance.

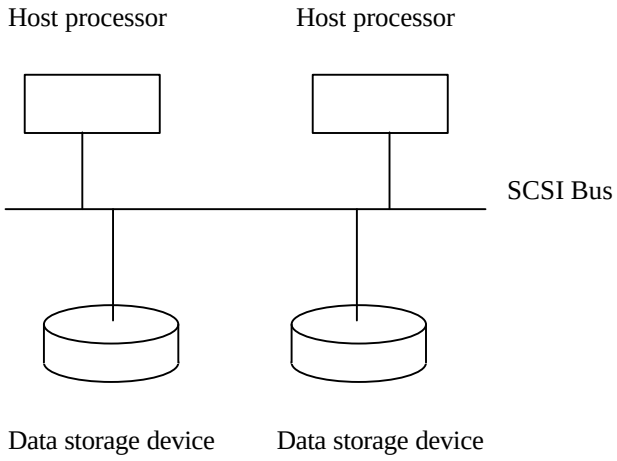
If the High Availability system supports an active standby mode of operation, the multiple hosts and multiple storage devices may be actively shared. Two or more host processors may be both performing I/O operations to a single disk. Coordination between the host processors is needed to prevent data corruption when using this mode of operation. This coordination may be done using communication over the SCSI bus or by an alternate path such as an ethernet connection between the multiple hosts.

There are a number of system issues that must be considered to achieve a High Availability system, including the failover of network traffic, shared access control, and management of the many issues related to system

security, device naming, and shared device access control. These issues are not directly related to the use of the SCSI bus and are beyond the scope of this Technical Report.

The general configuration of this type of system is shown in Figure 1. (The location of devices on the SCSI Bus as shown in this figure is for illustrative purposes only. No requirement about device location is to be inferred from the figure.)

Figure 1 SCSI-3 Multi-Host Configuration



This approach to building High Availability system has a weakness in that the SCSI bus itself connects all the components in the system and is therefore a single point of failure. However, since the bus is a passive component it is considered very reliable. Clearly in such a system every possible effort should be made to ensure that connectors remain securely attached, that cables be routed so as to not be tripped over, etc.

Note that if active SCSI bus terminators are used, or if active circuits are used to extend the length of the SCSI bus, these active devices must be accounted for in the High Availability strategy.

This Technical Report recognizes that the requirements for a device operating as a host are slightly different from those of a device operating as a target, and that diagnostic software has a slightly different set of requirements from the boot software or the run-time driver. Separate clauses in the text are used to describe these and other special cases.

The basic requirement for High Availability systems that are constructed using Multi-Host SCSI is that the hosts and devices must be capable of coexisting on a SCSI bus without interference, even in the case when a host or device fails.

This requirement is met by ensuring that at the highest level the system provides

- support for Multi-Host operation,
- support for hot plugging of all components except the SCSI bus itself,
- high reliability at the component hardware level as a result of good design, and
- good system-wide responsiveness to failure situations.

6. Fundamental Requirements

The SCSI system, including all devices, components, hardware and software, shall conform with all mandatory requirements of the latest revision of the SCSI-3 standard extant at the time of manufacture.

Because the standards that describe SCSI-3 are updated on independent schedules it may be difficult for a device to identify with full detail the exact version number for every standard applicable in a given situation. It

is recommended that device vendors identify their product as conforming to the versions of the applicable SCSI standards that were current as of the date when the design was finalized. By using this approach a customer may identify the state of the SCSI-3 standard as it applies to the device in question.

In particular, the SCSI system 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 Technical Report.

Specific additional exclusions and expansions to the SCSI-3 standard are described in this Technical Report.

7. SCSI-3 System Level Requirements

This clause describes system-wide requirements for Multi-Host High Availability configurations. Many of the requirements mentioned in this clause are expanded in later clauses.

7.1 Generic SCSI-3 System Level Requirements

The basic system-level requirement is to support hot plugging of both SCSI devices and SCSI hosts in a Multi-Host environment.

Mode page setting, logical unit reservations, ID assignments, and the use of bus resets must all be coordinated between hosts

7.2 Specific SCSI-3 System Level Requirements

7.2.1 Configuration Rules and Recommendations

It is recommended that the use of "Y" cables and the single-connector option be chosen for all designs. See Figure 3 below. This approach maximizes configuration flexibility and provides the opportunity to maximize system availability. Even a low-end system that is primarily intended for single host applications may eventually be added to a high availability system, and therefore should implement the single connector option.

Because of the SCSI bus stub length restrictions described below, the "Y" cable approach is usually applicable only to those situations where the enclosure contains devices that make separate individual external connections.

- For example, a storage subsystem with a RAID controller might have only one external connection to the RAID controller.
- For example, a host computer might only have a single host adapter that makes an external connection.
- For example, a storage subsystem with a pair of RAID controllers might have two external connectors, each intended for use by one controller by itself.
- For example, a storage subsystem consisting of a number of disks on one SCSI bus will probably not use the "Y" cable approach because it is too difficult to meet the stub length requirements.

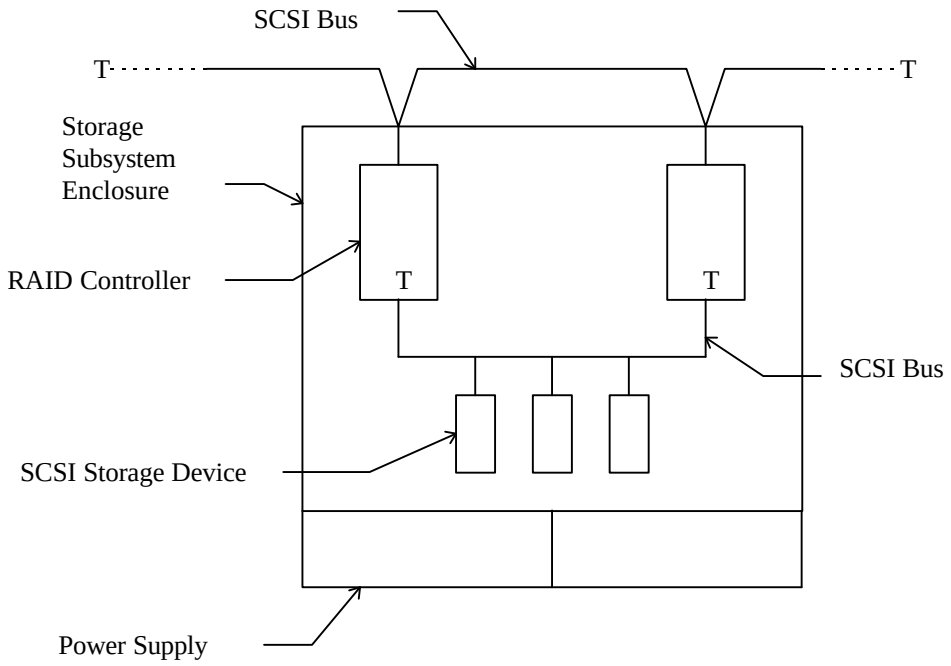
It is recommended that the mass storage devices in a High Availability system be housed in a self-contained enclosure that is separate from the host enclosures. By providing redundant power supplies and cabinet services a storage subsystem can provide SCSI service to more than one host system, with minimal disruption if one host system malfunctions.

If the redundant storage devices are all contained within a single enclosure, it is recommended that dual redundant independent power supplies be provided 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 enclosure then this is not as important.

For example, a High Availability storage subsystem enclosure might contain a pair of RAID controllers each with a single external SCSI connection, a number of internal SCSI disk drives connected to both controllers, and two power supplies to provide redundancy. The RAID software in combination with the redundant hardware provides data availability. A configuration of this type is shown in Figure 2.

Note that in Figure 2 the SCSI bus used to connect the hosts to the subsystem is separate from the SCSI bus used inside the subsystem to connect the controller to the devices. Each of these SCSI busses has its own termination, termpower, and configuration domain.

Figure 2. Typical Storage Subsystem Enclosure



7.2.2 Resource Sharing Rules Applicable to All Devices

All High Availability SCSI devices shall implement at least the "SCAM tolerant" level of SCAM as described in SPI Annex B. Full SCAM support is recommended.

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.

- 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.
- 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 affect active I/O on the bus.

High Availability Multi-Host systems shall have a mechanism to coordinate the access to shared data. If this coordination is done using the SCSI bus as a communication method, the RESERVE and RELEASE commands shall be used to protect the shared data.[why not specify use of PERSISTENT RESERVE and release?]

High Availability Multi-Host systems shall have a mechanism to coordinate the mode page settings on shared devices. The details of mode page coordination are vendor specific. [goal: remove all optional behavior]

Since the status of a reservation may change upon removal of device power, the hosts shall coordinate the reservations between themselves. The persistent reservation option may be used to improve this coordination. [goal: remove all optional behavior]

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

Device mode pages shall 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 shall 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.

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

8. SCSI-3 Physical Requirements

This clause describes the requirements placed on physical SCSI components in order to support High Availability.

8.1 Generic SCSI-3 Physical Requirements

The basic physical requirement is that the SCSI bus should be capable of supporting multiple hosts in a dynamic configuration. This means that it should be possible to construct a bus configuration that includes multiple hosts, storage devices, subsystems or other components such that changes can be made to the configuration without the need to perform objectionable levels of hardware manipulation. For example, it should not be necessary to remove external enclosure panels in order to obtain access to a connector needed to perform a device hot plug operation.

For most installations the SCSI bus should be expected to be at least several meters in length. This is because the use of multiple hosts and multiple storage devices generally implies the use of multiple enclosures, which in turn implies that the interconnecting bus is of some significant length. This may in many cases also imply that the use of differential SCSI-3 is preferred, although this is not required.

The installation should protect against invalid SCSI bus configurations. It is possible to assemble valid SCSI-3 components into systems that don't work by using an improper combination of available hardware. It is desirable that the system have a means of performing an automatic verification of the configuration to avoid this situation. 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. Since the SCSI protocol does not support the discovery of invalid configurations, the system should have a method for restricting host adapters so that they will not negotiate wide or fast methods of operation (i.e. will always operation in narrow synchronous mode). Control of this behavior should be under the manual control of the system administrator, since automatic methods of backing down to a lower level of performance may mask errors resulting from conditions not related to the configuration in a properly configured system.

The SCSI bus should support "device plugging" (the removal or insertion of a device on the SCSI bus) of devices or hosts without disturbing the other devices on the bus.

8.2 Specific SCSI-3 Physical Requirements

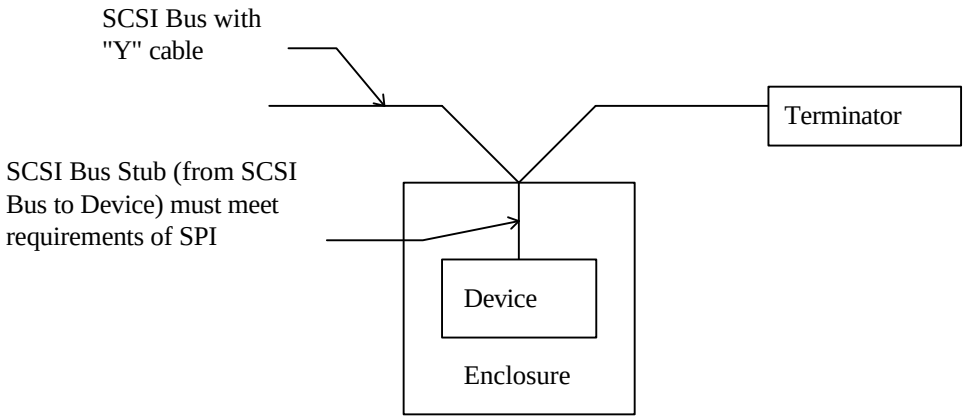
Device enclosures should be designed so that they need not be opened to change the configuration from the single-host configuration to the Multi-Host configuration. This allows newly delivered enclosures to be used in either environment without the need for a physical configuration process during installation.

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. To accomplish this requirement, SCSI devices, host adapters, and enclosures shall conform to one of the two cabling options listed below. 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.

8.2.1 Single Connector Option

The preferred option is the single 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, external to all enclosures. The stub length of the connector and cable inside the enclosure shall meet the requirements of SPI clauses 6.4 and 6.5.

Figure 3. Single Connector Option

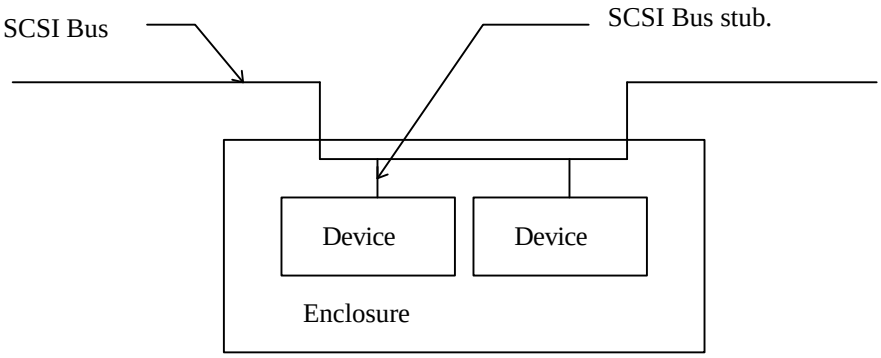


8.2.2 Double Connector Option

A less desirable option is the double connector option. The enclosure shall be implemented with two external SCSI connectors. In this case the SCSI bus enters and exits the enclosure 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 clause 5.2.

It should be noted that while enclosures that use the two connector option allow more internal wiring flexibility, they 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 maintains internal bus continuity when devices are removed from the enclosure.

Figure 4. Double Connector Option



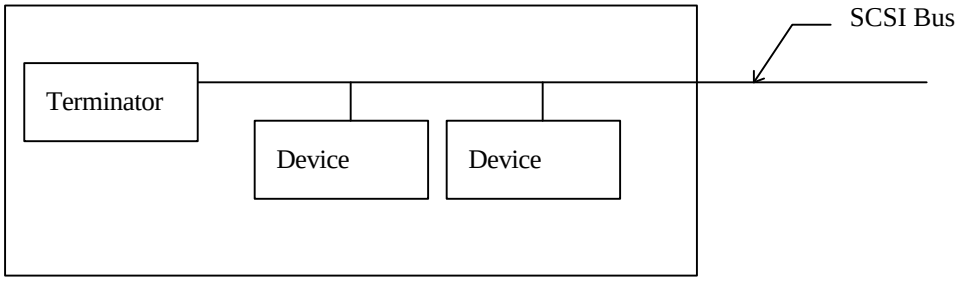
8.2.3 Enclosure with Internal Hosts and Storage

Many small desktop computer systems are used with a combination of internal and external storage. For example, the system disk may be an internal disk while the data storage disks and backup devices might be in an external enclosure. It is desirable to construct such systems in a way such that they may be used in High Availability environments. In either case the SCSI bus must be terminated properly.

If such a system is designed for use in single user environments only, the physical position of the SCSI bus terminators is not important. However, if it is to be used in a High Availability environment then it must be possible to gain access to the terminators without disassembling the enclosure. The best way to accomplish this is to use the double connector option with both SCSI bus terminators mounted externally. It is also acceptable to use the single connector option, but this can be difficult because of the stub length restriction.

All other configurations are unsuitable for use in High Availability configurations and must be avoided. Specifically, it is undesirable to have an enclosure with a single external connector with a SCSI bus terminator embedded in the enclosure.

Figure 5. Wrong Way to Build Enclosure



8.2.4 Cable Requirements

The SCSI cable plant shall meet the recommendations of SPI Annexes D and F. Because this Technical Report is intended to maximize the interoperability of components that are designed for use in High Availability systems, and since the SPI Annexes describe a set of requirements that can be expected to support satisfactory operation within certain configuration rules, a specific set of cable length restrictions is useful and necessary.

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.

Cable lengths for a complete system shall not exceed the following values.

8.2.4.1 Single Ended Cable Length Requirements

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

8.2.4.2 Differential Cable Length Requirements

All speeds: 25 meters

8.2.4.3 Low Voltage Differential Cable Length Requirements

tbd

8.2.4.4 Backplane Length Requirements

tbd

8.3 Device Plugging

SPI describes four cases in which devices may be removed or inserted from a SCSI bus. For each case the complete applicable requirements are listed. Refer to SPI Annex A. These cases are summarized as follows:

- Case 1 - Power off during removal or insertion.
- Case 2 - RST signal asserted continuously during removal or insertion.
- Case 3 - Current I/O process not allowed during insertion or removal.
- Case 4 - Current I/O process allowed during insertion or removal.

For High Availability systems it is desireable that the operation of removing or inserting a device on the bus cause the smallest possible disruption of bus traffic. Case 1 clearly causes substantial disruption at the system level. Cases 2 and 3 cause less system disruption but because the SCSI bus is stalled for the duration of the removal or insertion operation, they are considered unacceptable for High Availability systems.

High Availability systems shall support Case 4 device removal and insertion. This is referred to as "hot plugging" in the remainder of this Technical Report.

9. SCSI-3 Electrical Requirements

This clause describes the electrical requirements applicable to all devices connected to the SCSI bus.

9.1 Generic SCSI-3 Electrical Requirements

Since the primary goal of High Availability systems is to be able to maintain continuous operation in the event of a component failure, a basic requirement of such systems is that component replacement should be possible while the system is in operation. To support this, the SCSI bus and associated electronics should support hot plugging. Support for hot plugging implies some additional requirements, as follows. Refer to SPI for the specific electrical requirements for supporting hot plugging.

After a component is replaced using a hot plugging operation, power must be re-applied to the component. The component must power-up in a fashion that does not disrupt traffic on the SCSI bus.

Bus termination needs to be external to the enclosure. This allows the enclosure to be removed in the event that it needs to be replaced.

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.

Devices should be tolerant of electrostatic discharge on external connector pins.

9.2 Specific SCSI-3 Electrical Requirements

9.2.1 Termination

All SCSI bus termination devices shall be mounted externally to all enclosures. A device shall not be terminated internally within an enclosure. A device shall not be 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.

[The following assertion is untrue because a bus extender connects two physical busses into one logical bus. It could say something like: "If a bus extender or converter is located at the end of a bus segment, ...] Since bus extenders and converters (e.g. single-ended to differential) are always located at the end of an electrical bus segment, bus termination shall be provided nearby. This termination shall be external to the extender or converter.

Terminator power shall be supplied as described in SPI clause 7.3, except that "optional internal terminators" shall not be used. Redundant sources of terminator power shall be supplied.

9.2.2 Power Cycling

The driver and receiver electronics should be well-behaved during power cycles. Glitches or other irregular signals shall not be caused on the bus as a result of the application or removal of power to the device during or in preparation for the plugging operation. The bus drivers , receivers, and all other electrical connections to the data, REQ, ACK, and control lines on the device shall maintain the high-impedance state during power-up cycles until the drivers are enabled and during power-down cycles after the drivers have been disabled. Refer to SPI clauses 7.1.2 and 7.2.2.

After power is initially applied to a device after it is plugged, the unit attention flag shall be set for each initiator on each valid logical unit.

9.2.3 Hot Plugging

In order to ensure 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 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.

10. SCSI-3 Logical and Command Requirements

This clause describes how SCSI-3 messages and commands are used in a High Availability environment.

10.1 Generic SCSI-3 Logical and Command Requirements

The SCSI bus reset mechanism is not well suited for use in a Multi-Host environment because it is extremely disruptive to in-progress I/O operations. In a Multi-Host environment targets must maintain some data structures on a per-initiator basis, and must perform message phase actions so as to cooperate with other devices on the bus.

10.2 Specific SCSI-3 Logical and Command Requirements

10.2.1 Device Identification

The device ID mechanism used by the bus should be fully supported by the host's operating system. 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 should 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, High Availability SCSI devices shall support the vital product data unit serial number page.

Every device shall have a unique identification string, consisting of the vendor name and model name from the standard Inquiry data, concatenated with the device's serial number. If the device implements SCAM, the SCAM identifier string will serve this purpose.

10.2.2 Resets

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. SCSI Targets shall not assert the SCSI RST signal.

Since specific device resets (BUS DEVICE RESET) may interfere with device activity that was started by another host, device resets shall be sent only by hosts. Hosts that issue BUS DEVICE RESETs shall coordinate their use between themselves.

10.2.3 Renegotiation of Bus Options

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.

Hot plugging operations shall be done in such a way that any required renegotiations are performed.

All initiators should renegotiate any bus options such as wide or synchronous before issuing any command resulting in data transfer, including the INQUIRY or REQUEST SENSE commands. Refer to SIP clauses 8.2.12 and 8.2.15 for additional rules.

Host adapter devices shall offer a mode of operation in which the adapter does not negotiate for anything higher than narrow synchronous operation. This maximizes the chance that the system will work even in the

case of invalid cable configurations such as mixed wide and narrow cables and all wide devices. (Note that this does not help in configurations where the maximum bus length is exceeded.)

Upon host bootup, the console software of High Availability SCSI initiators shall negotiate data transfer width (using the WDTR message) and synchronous data transfer speed and offset (using the SDTR message) before attempting to execute any SCSI command that requires a data-in or data-out phase (including the INQUIRY command) to any target. Initiators in bootup mode 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.

10.2.4 Message and Command Interpreter Validity

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

Devices shall implement all the mandatory SCSI-3 commands for the device type they report. Optional commands that are not implemented should be handled properly according to the SCSI-3 standard.

Devices shall power up and complete their self tests properly regardless of the state of the SCSI bus. This includes cases of no terminator power, no termination, no bus attached, activity on an attached bus, and reset asserted. Devices that do not meet these requirements may experience difficulty if the devices on the bus do not power up simultaneously.

High Availability SCSI devices shall return a status of CHECK CONDITION with sense key of ILLEGAL REQUEST for any unsupported command. [difficulty here of what is definition of "device"--does it include initiators?]

High Availability SCSI devices shall return a status of CHECK CONDITION with sense key of ABORTED COMMAND/MESSAGE ERROR for any unsupported message. [How many Bytes of an unsupported message should the target swallow before giving up?]

Parity checking shall be enabled. Parity shall be checked on all received data, in any information transfer phase and during the Selection or Reselection phases. Invalid parity shall be handled in accordance with the SCSI-3 standard.

10.2.5 Per-Initiator Data Structures

The SCSI unit attention flag shall be maintained on a per-initiator basis in the device, as required by SAM in paragraph 5.6.5. The unit attention condition shall be generated for each initiator on each valid logical unit whenever the target receives a BUS DEVICE RESET message, a bus reset, or a power cycle. The unit attention condition shall persist on each logical unit for each initiator until that initiator clears the condition on each logical unit.

Targets shall be capable of maintaining separate sense data for each initiator and shall not return BUSY status to any initiator as a result of a pending contingent allegiance condition with any other initiator.

10.2.6 Handling ABORT Message

If an ABORT message is received, the device shall abort the current operation in a manner that does not cause loss or corruption of data. Parity errors shall not be written to the media, and data for which a GOOD status has been returned to the initiator shall be written to the media before processing the ABORT. Targets that require additional time for this buffer flushing operation shall return BUSY status in response to connection attempts. [Question of handling of case where device writes to media before returning status.]

11. SCSI-3 Target Device Requirements

This clause applies to each SCSI-3 device that is operating as a target.

11.1 Generic SCSI-3 Target Device Requirements

[what about linked commands?]

Devices shall properly handle bus resets that may occur at any time. Setup information shall 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 shall properly support simultaneous hosts. Devices shall 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. Sense data should be retained on a per-initiator basis.

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

Devices shall properly handle device reservation (using either RESERVE or PERSISTENT RESERVE) in a Multi-Host environment. Note that some systems may never issue RESERVE/RELEASE commands. However, since RESERVE/RELEASE provide a convenient mechanism for low level synchronization it is desireable that they be supported by all targets.

Devices shall support tagged commands in a timely manner. Adherence to this requirement has a substantial impact on the overall responsiveness of a Multi-Host system. Proper behavior is as follows:

1. The device may defer a given command until all the previously queued commands complete, regardless of the time those commands require. (For example, a FORMAT or REWIND command may take considerable time.)
2. The device may reorder commands (within the rules of tagged command queueing) to defer the execution of a given command until a later time.
3. From the time at which a given command could first have been executed, the device shall not defer execution of the command more than one second.

11.2 Specific SCSI-3 Target Device Requirements

11.2.1 Reset Support

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

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.

11.2.2 Initiator Support

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.

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, as stored in the unit attention flag, shall indicate whether 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-3 Unit Attention Condition conditions apply. Refer to SAM clause 5.6.5.

Devices shall either retain sense data on a per-initiator basis and also return the sense data to the correct initiator or shall stop processing during the auto contingent allegiance condition. Refer to SAM clause 5.6.1.

11.2.3 Logical Unit Support

If a target device supports multiple logical units, then mode pages shall be maintained on a per-logical unit basis.

High Availability SCSI target devices shall support Tagged Command Queuing. Note that sCSI-2 clause 7.8.2 requires that all comman d received with Simple Queue Tag message prior to a command received with Ordered Queue Tag message, regardless of Inititator, shall be executed before that command with the Ordered Queue Tag Message. This is essential for the correct operation of the queueing algorithm when used in Multi-Host systems.

High Availability SCSI target devices shall support drive-based Bad Block Replacement (BBR) as described in SCSI-3. This is required in Multi-Host systems to avoid potential wastage of revectoring resources in the case where two hosts attempt simultaneous revectoring.

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. In order to prevent retries from timing out other devices, High Availability SCSI devices shall delay at least 2.4 milliseconds between retry attempts. Targets shall respond to Initiator selection attempts that occur during the 2.4 millisecond delay between retry attempts.

A target having multiple queued commands for an initiator that fails to respond to reselection, including retries, shall abort all commands from that initiator in the queue. The device shall generate a contingent allegiance conditions for the timed-out initiator with a sense key of HARDWARE ERROR and an ASC/ASCQ of SELECT OR RESELECT FAILURE. This allows multi-initiator environments to continue operation with minimal impact. After having aborted all commands for the timed-out initiator, the device shall generate a contingent allegiance condition for the timed-out initiator with a sense key of HARDWARE ERROR (04h) and an ASC/ASCQ value of SELECT OR RESET FAILURE (45/00h).

Target devices that do not support wide SCSI shall respond to initiator attempts to negotiate wide operation by returning the WDTR message in sequence specifying narrow operation rather than sending MESSAGE REJECT. Completing the WDTR sequence rather than rejecting it resets any previously negotiated synchronous data transfer agreement to the default asynchronous mode. This will prevent bus hang conditions due to synchronous/asynchronous mismatch between targets and initiators. Targets shall track the negotiated wide transfer agreements on a per-initiator basis.

In every case where a WDTR message is sent, it should be followed by an SDTR. This method guarantees that the host and device are in agreement with respect to these two modes of operation.

11.2.4 Error Condition Support

Targets shall manage fault conditions by going to the Bus Free state only in those cases where this action is required by SCSI-3 for catastrophic errors. Exception conditions other than those requiring Bus Free per SCSI-3 shall be handled by other means such as going to STATUS phase with a check condition, by sending a message reject, or by retrying the phase as appropriate.

A target device that terminates a WRITE command with a check condition due to parity errors shall not write the associated data to the media. [Question of cache contents.]

11.2.5 RESERVE/RELEASE Support

Targets shall support the RESERVE and RELEASE SCSI commands. These commands allows hosts to allocate devices with exclusive access.

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 (Prevent bit cleared to 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.

12. SCSI-3 Initiator Device Requirements

This clause describes the requirements that apply to SCSI-3 devices that are operating as initiators.

12.1 Generic SCSI-3 Initiator Device Requirements

The High Availability configuration rules must be followed.

The SCSI bus is expected to be active at all times, so every software and hardware function must operate correctly even during power cycles and other system-level state changes.

Hosts should minimize the number of bus reset operations that they initiate. This means that a host should attempt to avoid the reset condition during initialization, normal processing, and shutdown. A bus reset should be asserted 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.

Logical units must ensure that adequate command queue space is reserved for cases where multiple initiators wish to communicate at the same time.

12.2 Specific SCSI-3 Initiator Device Requirements

12.2.1 Configuration Rules

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 may be limited to two hosts that cannot be hot-plugged. This is not adequate for a High Availability system.

The host shall implement Target Mode operation as a processor device, and shall support all mandatory requirements of SCSI-3 pertaining to the processor device type.

12.2.2 Bus Support At All Times

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 software, the host adapter sooftware, 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. In particular, if the Initiator has a connection active, that connection shall be ended either by following the Target's phase and data transfer requests until the next Bus Free condition, or by asserting ATN and sending the ABORT message. If the second alternative is followed, the Initiator must still react to the target until the Bus Free condition.

Even 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: [goal: remove all optional behavior]

- Disable selection as a target in the console or adapter card. This option is most appropriate for adapters that have little intelligence on them.
- 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.
- Enable SELECTIONs and return SCSI status of BUSY, and then return the COMMAND COMPLETE message.

12.2.3 Reset Support

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.

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.

From the viewpoint of a given initiator the bus may be hung in communication either with that initiator or with another initiator. In either case, the initiator shall use the same procedure to attempt recovery. This approach is taken because otherwise the bus hang recovery algorithm is dependent on the inter-host coordination method.

The initiator that suspects that a target is hung should first 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 by issuing a Clear Queue message, 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.

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 should 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.)

When a device receives a hard reset, it shall first ensure that all cached data for which good status has been returned is written to non-volatile media prior to processing the reset. If the device is a sequential access device it shall additionally write an EOD to the medium after flushing the cached data, then it shall rewind the media to BOT.

For sequential devices, it is acceptable to return BUSY status while the device is flushing its buffer and rewind operations are complete. The preferred action is to immediately process the command. Accepting the command and disconnecting to wait until the device can process it will result in host timeouts.

If a device receives multiple valid bus resets in succession, it shall process the reset and recover within 250 msec. [need clarification here again re "device" vs "device" as used above in 60 second discussion]

12.2.4 Command Queue Support

Initiators shall not use all the tag queue depth in a device. [does there need to be a standard way to determine the queue tag depth?]

The initiator shall reserve some number of tag queue elements so that other initiators may still send commands (such as INQUIRY) to the device while it is in use by another initiator.

13. SCSI-3 Requirements for Specific Device Types

Certain additional requirements beyond those required in SCSI-3 may be needed in High Availability systems.

13.1 Direct Access Device Type Requirements

TBD

13.2 Sequential Access Device Type Requirements

TBD

13.3 Other Device Type Requirements

TBD

14. Effect On Existing Standards

No changes are required in the existing SCSI-3 standard to support Multi-Host High Availability systems. This Technical Report describes a number of additional restrictions and implementation rules that, when applied in addition to the requirements of the SCSI-3 standard, allow systems to be constructed that have a high degree of tolerance to SCSI component failures.

Future versions of the SCSI-3 standard may be modified to include some of the rules described herein.

Many of the concepts described in this Technical Report apply to interconnects described in other ANSI standards such as Fibre Channel and SSA. Similar reports applicable to those standards may be appropriate in the future.