digital

X3T10/96-263, Revision 0

To: X3T10 Committee

From: Tom Coughlan
Digital Equipment Corporation
Mail Stop ZKO-3-4/U14
110 Spitbrook Road
Nashua, New Hampshire
Telephone: 603-881-0933
E-mail: tom.coughlan@zko.mts.dec.com

Subject: Proposal for an Additional Persistent Reservation Type

Summary

This proposal adds one new reservation type to the Persistent Reservation management method, described in SPC. This reservation type is required by systems that allow multiple initiators to simultaneously access a SCSI logical unit.

Revision History

This proposal was first discussed on the SCSI Reflector, starting with a message from me on September 4, 1996. The concept was also presented at the SCSI Working Group on September 9, 1996. The working group had a favorable response and asked for a complete proposal, for final review and approval.
There have been three changes since the original proposal on the reflector:

1. The proposed mechanism to allow a group of initiators to be preempted with a single command has been removed, because it adds unnecessary complexity.
2. The proposal for a “shared access, registrants only” reservation type has been replaced by a “write exclusive, registrants only” reservation. This change allows non-registered initiators to bootstrap in a read-only mode, then perform a registration before writing.
3. A clarification of the existing SPC text has been added. The new text stipulates that the Preempt action, and the Preempt and Clear action shall execute without error even if there is no reservation to preempt. In particular, the Clear part of the action shall still occur.

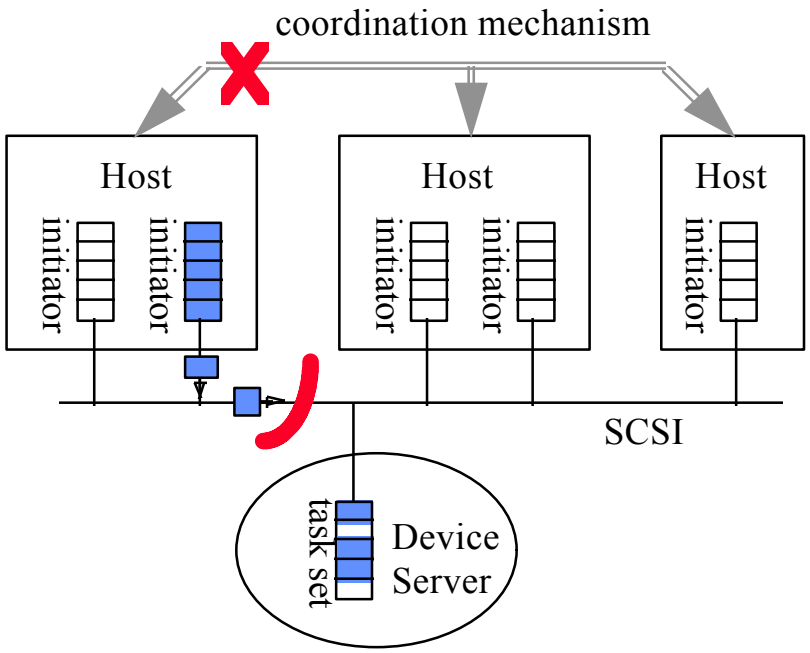
The Need for this Change

Initiators in a multiple-initiator SCSI system must coordinate their access to shared logical units. Generally, when a member of such a system becomes unresponsive, some amount of clean-up must be done before the surviving members can continue. In SCSI, this requires determining that:

1. There are no tasks that modify the media queued to the shared device server by the unresponsive initiator.
2. The device server does not have an ACA condition with the unresponsive initiator.

3. The device server does not have a reservation from the unresponsive initiator.
4. Subsequent commands that modify the media from the unresponsive initiator will not be accepted by the device server. (For example, commands that are in-flight when an initiator halts, or that may subsequently emanate from an intelligent adapter.)

The following diagram serves to illustrate the problem:



Since multiple-initiator systems are often used to achieve high availability, it is important to be able to remove a failed node quickly, and to minimize the disruption of the surviving initiators.

The current Persistent Reservation management method provides the necessary features to solve this problem for systems in which just one surviving initiator has access to a shared device after an unresponsive initiator is removed from the system. An additional provision is needed for multiple-initiator systems that allow multiple initiators to simultaneously access the shared storage device.

The Proposal

This proposal adds a new Persistent Reservation type, called “Exclusive Write, Registrants Only”. This reservation allows all initiators to read the device, but only registered initiators can write to the device. This allows an initiator to do read-only bootstrap from a shared device, then, once coordination with the other accessors is achieved, register, possibly issue an Exclusive Write, Registrants Only reservation, then issues write commands. When an initiator becomes unresponsive, one or more of the survivors issues a Preempt and Clear. This action removes all state associated with the unresponsive initiator from the device server, and blocks subsequent writes. When the affected initiator recovers, it shall ensure that its old commands are cleaned-up, then re-register and re-reserve.

Specific Changes

The following changes are needed. (This description is based on SPC Rev. 10.)

Section 7.12.3.2 Persistent Reservations Type

Add the following to Table 41.

Table 41 - Persistent Reservation Type Codes

Code	Name	Description
5h	Write Exclusive, Registrants Only	Reads Shared: Any application client on any initiator may execute commands that perform transfers from the storage medium or cache of the logical unit to the initiator. Writes Exclusive: Any command that performs a transfer from an initiator that has not previously performed a Register service action with the device server, to the storage medium or cache of the logical unit, shall result in a reservation conflict. Additional Reservations Allowed: Any initiator may reserve the logical unit or extents or elements as long as the persistent reservations do not conflict with any reservations that are already known to the device server. See table 42.
6-Fh	Reserved	

Change Table 42 as follows:

Table 42

Persistent Reservation That is Being Attempted		Persistent Reservation That Is Held											
		Read Shared		Write Exclusive		Read Exclusive		Exclusive Access*		Shared Access*		Write Exclusive, Registrants Only	
		LU	EX	LU	EX	LU	EX	LU	EX	LU	EX	LU	EX
Read Shared	LU	N	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y
	EX	N	N	Y	O	Y	O	Y	O	N	N	Y	O
Write Exclusive	LU	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	EX	Y	O	Y	O	Y	O	Y	O	Y	O	Y	O
Read Exclusive	LU	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	EX	Y	O	Y	O	Y	O	Y	O	Y	O	Y	O
Exclusive Access*	LU	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	EX	Y	O	Y	O	Y	O	Y	O	Y	O	Y	O
Shared Access*	LU	N	N	Y	Y	Y	Y	Y	Y	N	N	Y	Y
	EX	N	N	Y	O	Y	O	Y	O	N	N	Y	O
Write Exclusive, Registrants Only	LU	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
	EX	Y	O	Y	O	Y	O	Y	O	Y	O	N	N

Key:

LU = Logical Unit scope
EX = Extent or Element scope
* = Conflicts with all reservation requests from other initiators

N = no conflict
Y = conflict
O = conflict occurs if extent or element overlaps with existing extent or element reservation

Section 7.13.1.5 Preempt

Add the following text, which is similar to the text in Section 7.13.1.3, pertaining to the Release action.:

It shall not be an error to send a PERSISTENT RESERVE OUT specifying a Preempt service action when no persistent reservation exists for the initiator identified by the Service Action Reservation key.

Section 7.13.1.6 Preempt and Clear

Add the following text, which is similar to the text in Section 7.13.1.3, pertaining to the Release action.:

It shall not be an error to send a PERSISTENT RESERVE OUT specifying a Preempt and Clear service action when no persistent reservation exists for the initiator identified by the Service Action Reservation key. Furthermore, if the key is registered, the Clear portion of the action shall execute normally.