Date: November 4, 2003

To: T10 Committee (SCSI)

From: Roger Cummings (VERITAS)

Subject: T10/03-289r1 Only If Reserved Proposal

Revision History

03-289r0 (August 27, 2003) First Revision
03-289r1 (November 4, 2003) Clarified the specific command to which the OIR bit applies.

Related Documents

03-231r0 (July 2, 2003) Two persistent reservations problems - latest information. Feedback on earlier approaches and results from latest testing & investigation.
spc3r15 – SCSI Primary Commands – 3 revision 15

Backgound

This proposal has arisen from recent research into persistent reservations, but it is a separate proposal which stands alone from that work. The feature proposed here will work with (original) reservations, or any type of persistent reservation, and will directly address some significant problems encountered by designers of backup and restore applications. Adoption of this proposal will allow simplification of existing code, and provide much-improved resistance against a particularly destructive class of data corruption errors that have been encountered in a number of situations in the field.

Overview

Reserves and Releases are presently used extensively in situations where sequential-access devices are employed. There are known to be situations where multiple reserves and releases are issued against the same device by applications, device drivers and other entities that coexist in a computer system.

These situations occur because there is no way to test that an (original) reservation is still in existence for an I_T nexus. Issuing a RESERVE command returns the same status if a reservation to the issuing I_T nexus existed before the command was received or not. Similarly, issuing a RELEASE command returns the same status if a reservation to the issuing I_T nexus existed before the command was received or not.

Designers of backup and restore applications are very concerned to avoid writing to unreserved serial access devices, because they can be subject to later and almost undetectable data corruption resulting from accesses from other Initiators (either servers or data copy engines). The combination of this concern and the lack of a way of testing a reservation results in the profligate use of reserves described above.

However the provision of a test facility would not by itself solve this problem, because by definition there will always be a time interval between the completion of the test and the issuance of a data transfer command. To avoid this “hole” what is needed is a way to ensure that a command will only be executed by the device server if a reservation exists for the I_T nexus under which the command is issued. If no such reservation exists, the command would not be performed, and the device server would return a new status to indicate that the device server is not reserved.

Proposal

The document proposes the definition of a new bit for the Control mode page, called Only If Reserved (OIR). The definition of this bit has been extended from the background given above to also cover persistent reservations.

Suggested Changes

7.4.6 Control mode page

The Control mode page (see table 223) provides controls over several SCSI features that are applicable to all device types such as tagged queuing and error logging.

Table 223 — Control mode page

Bit Byte	7	6	5	4	3	2	1	0
0	PS	SPF (0b)	PAGE CODE (0Ah)					
1	PAGE LENGTH (0Ah)							
2	TST			Reserved		D_SENSE	GLTSD	RLEC
3	QUEUE ALGORITHM MODIFIER				Reserved	QERR		Obsolete
4	TAS	RAC	UA_INTLCK_CTRL		SWP	Obsolete		
5	OIR	Reserved				AUTOLOAD MODE		
6	Obsolete							
7								
8	(MSB)	BUSY TIMEOUT PERIOD						
9								(LSB)
10	(MSB)	EXTENDED SELF-TEST COMPLETION TIME						
11								(LSB)

A task set type field (TST) specifies the type of task set in the logical unit (see table 224).

Table 224 — Task set type

Value	Description
000b	The logical unit maintains one task set for all initiator ports
001b	The logical unit maintains separate task sets for each initiator port
010b - 111b	Reserved

If the logical unit maintains separate mode pages for each initiator port, the TST field, if changeable, shall reflect in the mode pages for all initiator ports the state selected by the most recent MODE SELECT from any initiator port. If the most recent MODE SELECT changes the setting of this field the device server shall establish a unit attention condition for all initiator ports except the one that issued the MODE SELECT command (see SAM-2). The device server shall set the additional sense code to MODE PARAMETERS CHANGED.

A descriptor format sense data (D_SENSE) bit set to zero indicates that the device server shall return the fixed format sense data (see 4.5.3) in the same I_T_L_Q nexus transaction (see 3.1.39) as a CHECK CONDITION status. A D_SENSE bit set to one indicates that the device server shall return descriptor format sense data (see 4.5.2) in the same I_T_L_Q nexus transaction as a CHECK CONDITION status.

A global logging target save disable (GLTSD) bit set to zero allows the SCSI target device to provide a vendor specific method for saving log parameters. A GLTSD bit set to one indicates that either the SCSI target device has disabled the vendor specific method for saving log parameters or, when set by the application client, specifies that the vendor specific method shall be disabled.

A report log exception condition (RLEC) bit set to one specifies that the device server shall report log exception conditions as described in 7.2.1. A RLEC bit set to zero specifies that the device server shall not report log exception conditions.

The QUEUE ALGORITHM MODIFIER field (see table 225) specifies restrictions on the algorithm used for reordering tasks having the SIMPLE task attribute (see SAM-3).

Table 225 — Queue algorithm modifier

Value	Description
0h	Restricted reordering
1h	Unrestricted reordering allowed
2h - 7h	Reserved
8h - Fh	Vendor specific

A value of zero in the QUEUE ALGORITHM MODIFIER field specifies that the device server shall order the processing sequence of tasks having the SIMPLE task attribute such that data integrity is maintained for that initiator port (i.e., if the transmission of new SCSI transport protocol requests is halted at any time, the final value of all data observable on the medium shall have exactly the same value as it would have if all the tasks had been given the ORDERED task attribute).

A value of one in the QUEUE ALGORITHM MODIFIER field specifies that the device server may reorder the processing sequence of tasks having the SIMPLE task attribute in any manner. Any data integrity exposures related to task sequence order shall be explicitly handled by the application client through the selection of appropriate commands and task attributes.

The queue error management (QERR) field (see table 226) specifies how the device server shall handle other tasks when one task receives a CHECK CONDITION status (see SAM-3). The task set type (see the TST field definition in this subclause) defines which other tasks are affected. If the TST field equals 000b, then all tasks from all initiator ports are affected. If the TST field equals 001b, then only tasks from the initiator port that receives the CHECK CONDITION status are affected.

Table 226 — Queue error management (QERR) field

Value	Definition
00b	If an ACA condition is established, the affected tasks in the task set shall resume after the ACA condition is cleared (see SAM-3). Otherwise, all tasks other than the task that received the CHECK CONDITION status shall be processed as if no error occurred.
01b	All the affected tasks in the task set shall be aborted when the CHECK CONDITION status is sent. If the TAS bit is set to zero, a unit attention condition (see SAM-2) shall be generated for each initiator port that had tasks aborted except for the initiator port to which the CHECK CONDITION status was sent. The device server shall set the additional sense code to COMMANDS CLEARED BY ANOTHER INITIATOR. If the TAS bit is set to, all affected tasks for initiator ports other than the initiator port for which the CHECK CONDITION status was sent shall be completed with a TASK ABORTED status and no unit attention shall be generated. For the initiator port to which the CHECK CONDITION status is sent, no status shall be sent for the tasks that are aborted.
10b	Reserved
11b	Affected tasks in the task set belonging to the initiator port to which a CHECK CONDITION status is sent shall be aborted when the status is sent.

A task aborted status (TAS) bit set to zero specifies that aborted tasks shall be terminated by the device server without any response to the application client. A TAS bit set to one specifies that tasks aborted by the actions of another initiator port shall be terminated with a TASK ABORTED status (see SAM-2).

The report a check (RAC) bit provides control of reporting long busy conditions or CHECK CONDITION status. A RAC bit set to one specifies that a CHECK CONDITION status should be reported rather than a long busy condition (e.g., longer than the busy timeout period). A RAC bit set to zero specifies that long busy conditions (e.g., busy condition during auto contingent allegiance) may be reported.

The unit attention interlocks control (UA_INTLCK_CTRL) field (see table 227) controls the clearing of unit attention conditions reported in the same I_T_L_Q nexus transaction (see 3.1.39) as a CHECK CONDITION status and whether returning a status of BUSY, TASK SET FULL or RESERVATION CONFLICT results in the establishment of a unit attention condition (see SAM-3).

Table 227 — Unit attention interlocks control (UA_INTLCK_CTRL) field

Value	Definition
00b	The logical unit shall clear any unit attention condition reported in the same I_T_L_Q nexus transaction as a CHECK CONDITION status and shall not establish a unit attention condition when a task is terminated with BUSY, TASK SET FULL, or RESERVATION CONFLICT status.
01b	Reserved
10b [a]	The logical unit shall not clear any unit attention condition reported in the same I_T_L_Q nexus transaction as a CHECK CONDITION status and shall not establish a unit attention condition when a task is terminated with BUSY, TASK SET FULL, or RESERVATION CONFLICT status.
11b [a]	The logical unit shall not clear any unit attention condition reported in the same I_T_L_Q nexus transaction as a CHECK CONDITION status and shall establish a unit attention condition for the initiator port that is the source of a task being terminated with BUSY, TASK SET FULL, or RESERVATION CONFLICT status. Depending on the status, the device server shall set the additional sense code to PREVIOUS BUSY STATUS, PREVIOUS TASK SET FULL STATUS, or PREVIOUS RESERVATION CONFLICT STATUS. Until it is cleared by a REQUEST SENSE command, a unit attention condition shall be established only once for a BUSY, TASK SET FULL, or RESERVATION CONFLICT status regardless to the number of commands terminated with one of those status values.

[a] A REQUEST SENSE command still clears any unit attention condition that it reports.

A software write protect (SWP) bit set to one specifies that the logical unit shall inhibit writing to the medium after writing all cached or buffered write data, if any. When SWP is one, all commands requiring writes to the medium shall return CHECK CONDITION status and shall set the sense key to DATA PROTECT and the additional sense code to WRITE PROTECTED. When SWP is one and the device type's command set defines a write protect (WP) bit in the DEVICE-SPECIFIC PARAMETER field in the mode parameter header, the WP bit shall be set to one for subsequent MODE SENSE commands. A SWP bit set to zero specifies that the logical unit may allow writing to the medium, depending on other write inhibit mechanisms implemented by the logical unit. When the SWP bit is set to zero, the value of the WP bit, if defined, is device type specific. For a list of commands affected by the SWP bit and details of the WP bit see the command standard (see 3.1.17) for the specific device type.

An only if reserved (OIR) bit set to one specifies that the device server shall only perform a command if a reservation or persistent reservation exists which allows access to the I_T nexus from which the command was received. When OIR is one and a command is received from an I_T nexus for which no reservations exists, the device server shall not perform the command. When OIR is one and a command is received from an I_T nexus for a logical unit or element upon which no reservation or persistent reservation exists, the device server shall terminate the command with CHECK CONDITION status, and shall set the sense key to NOT READY and the additional sense code to NOT RESERVED. Commands not affected by OIR shall be RESERVE, RELEASE, PERSISTENT RESERVE IN AND PERSISTENT RESERVE OUT. Commands affected by oir are defined with reference to the tables that define the commands allowed in the presence of various reservations in this standard (Table 31) and in the command standard (see 3.1.17) for the specific device type. Any command which has “Conflict” in any column of those tables shall be affected by OIR, except as noted above.

The AUTOLOAD MODE field specifies the action to be taken by a removable medium device server when a medium is inserted. For devices other than removable medium devices, this field is reserved. Table 228 shows the usage of the AUTOLOAD MODE field.

Table 228 — AUTOLOAD MODE field

Value	Definition
000b	Medium shall be loaded for full access.
001b	Medium shall be loaded for medium auxiliary memory access only.
010b	Medium shall not be loaded.
011b - 111b	Reserved

The BUSY TIMEOUT PERIOD field specifies the maximum time, in 100 milliseconds increments, that the application client allows for the device server to remain busy for unanticipated conditions that are not a routine part of commands from the application client. This value may be rounded down as defined in 5.4. A 0000h value in this field is undefined by this standard. An FFFFh value in this field is defined as an unlimited period.

The EXTENDED SELF-TEST COMPLETION TIME field contains advisory data that is the time in seconds that the device server requires to complete an extended self-test when the device server is not interrupted by subsequent commands and no errors occur during processing of the self-test. The application client should expect this time to increase significantly if other commands are sent to the logical unit while a self-test is in progress or if errors occur during execution of the self-test. Device servers supporting SELF-TEST CODE field values other than 000b for the SEND DIAGNOSTIC command (see 6.26) shall support the EXTENDED SELF-TEST COMPLETION TIME field.

Bits 0, 1, and 2 of byte 4 as well as bytes 6 and 7 provide controls for the obsolete asynchronous event reporting feature.

0.0.1 Table 28 and Annex C

Assign an additional ASCQ to indicate NOT RESERVED.