Sun Microsystems, Inc.
2550 Garcia Avenue
Mountain View, CA 94045
415 960-1300

X3T9.2/93-126

July 16, 1993

John Lohmeyer
Chairman, X3T9.2
1635 Aeroplaza Drive
Colorado Springs, CO 80916

Subject: Proposed architecture for configuring external RAID devices

The proposed RAID architecture (X3T9.2/93-003, revision 05) defines a series of models that move the Disk Array Conversion Layer (DACL) to various locations with respect to a host CPU and a group of disks. Those models that maintain the complete DACL mapping on the host side of the SCSI use configuration mechanisms that do not require any propagation of information across the SCSI. One model, the "Bridge Controller SCSI Disk Array" does require the configuration information to be exchanged across the SCSI.

I would like to propose a mapping of the configuration information into standard SCSI constructs using the MODE SELECT and MODE SENSE commands. This configuration information should be considered in the context of those SCSI addressing structures that allow a sufficient number of addresses to allow all required devices to be identified. For parallel SCSI, only 31 devices could be supported, since one of the 32 available addresses, the configuration address, would be required to not overlap the individual device addresses. Other SCSI protocols provide significantly deeper addressing structures. As an example, the addressing structure of FCP is 64 bits long, typically organized as a hierarchy of 4 16-bit addresses.

This proposal will be extended to meet any additional requirements.

Sincerely,

Robert N. Snively
Member, X3T9.2

RAID Configuration Commands

1 Introduction

1.1 Terminology

The DACL or Disk Array Conversion Layer provides Logical Unit Number and Logical Block mapping services from any addressable entity to the single drive or drives actually containing the referenced data.

The "RAID manager" is the addressable entity that contains the configuration management functions for the RAID subsystem. These configuration management functions provide the information that the DACL needs to properly manage the subsystem and store that information in the locations required by the DACL. A particular RAID subsystem may have multiple RAID managers at different entity addresses within the subsystem. As an example, a RAID manager may be defined that configures a group of grouped drives, each of which has its own RAID manager.

A "grouped drive" is an addressable entity that is a logical drive composed of one or more physical drives. The single-drive image is mapped by the DACL from the associated physical drives. The grouped drive may have any redundancy properties.

A "RAID drive" is an addressable entity that may be a grouped drive or a single drive. If the addressable entity is a single drive, the single drive behaves in the standard SCSI manner to all commands.

A "device" may be a RAID manager or any RAID drive, grouped or not.

1.2 Addressing

The Entity Address, defined and used by the SBP, GPP, and FCP documentation is used to address a device. The SIP documentation defines the Logical Unit Number for the same function.

The Entity Address allows the definition of a hierarchical address structure.

When the Entity Address is used as the destination of a SCSI command, the command applies to the addressed entity and may apply to the operation of all addressable entities below the addressed entity in that branch of the hierarchy.

When the Entity Address is used in a Mode Page, the address references information to be provided to or obtained from the addressed entity and may apply to all addressable entities below the addressed entity in that branch of the hierarchy.

2 SCSI commands used for RAID functions

2.1 MODE SELECT(6) command

The MODE SELECT(6) command is implemented as described in the applicable SCSI standard.

2.2 MODE SELECT(10) command

The MODE SELECT(10) command is implemented as described in the applicable SCSI standard. All RAID functions use pages that may exceed the size requirements of the MODE SELECT(6) command, so it is recommended that only the MODE SELECT(10) command be used for transmitting pages controlling the RAID functions.

2.3 MODE SENSE(6) command

The MODE SENSE(6) command is implemented as described in the applicable SCSI standard

2.4 MODE SENSE(10) command

The MODE SENSE(10) command is implemented as described in the applicable SCSI standard. All RAID functions use pages that may exceed the size requirements of the MODE SENSE(6) command, so it is recommended that only the MODE SENSE(10) command be used to receive pages describing the RAID functions.

3 Pages for managing RAID functions

The page code specifies which page or pages to return. Page code usage by the RAID manager is defined below.

Table 1: Page code usage for RAID

Page code	Description
TBD0	Disk drive group configuration and state page
TDB1	Internal operation mode page
TBD2	Drive state and information page
TDB3	Cached write configuration page
TBD4	Reserved
TBD7	

Page codes other than those listed above are defined by the applicable SCSI standard. Just as many fields are not implemented by individual drives, a grouped drive may elect to prohibit the presentation and setting of certain values that are not relevant or are managed automatically.

SCSI commands and pages addressed to the Entity Address of any drive are not examined or modified by entities higher along the path of the hierarchy.

The SCSI commands and pages addressed to the Entity Address of a grouped drive are processed according to the apparent disk properties provided by the grouped drive. As an example, a grouped drive composed of 4 drives striping full tracks together may provide the same number of records per track and cylinders per device that each individual drive implements, but may provide the appearance of four times the number of tracks per cylinder.

The the page code values should be selected to avoid conflicts with previously defined page codes.

3.1 Drive group configuration page

Table 2: Drive group configuration page

Bit Byte	7	6	5	4	3	2	1	0
0	PS	Rsvd	Page code (TBD0h)					
1	Reserved							
2	Page length							
3								
4	Reserved							
6								
7	Number of groups							
8	First group page							
n								
	. .							
x	Last group page							
x+(n-8)								

The Drive Group Configuration page is used in a MODE SELECT(10) command to create one or more drives composed of a group of individual physical drives. The MODE SENSE(10) command lists the composition of all grouped drives below the RAID manager addressed by the command. No information for individual drives is presented. The page above is valid only when the MODE SENSE/SELECT(10) command is addressed to the RAID manager.

PS

The parameters savable (PS) bit is unchanged from its definition in the applicable SCSI standard.

Number of groups

The Number of groups field contains the number of separate groups that are to be defined for the RAID manager or the number of separate groups that are managed by the RAID manager and are to be returned in this list.

Group page

The format of the Group page is defined in table 3.

Table 3: Group page format

Bit Byte	7	6	5	4	3	2	1	0
0	GROUP	BAD	Reserved					
1	Reserved							
2	Number of drives							
3								
4	Stripe block size							
7								
8	Group properties							
11								
12	Time stamp							
15								
16	Host world wide name							
23								
24	Peripheral world wide name							
31								
32	Reserved							
39								
40	First drive page							
63								
	. .							
x	Last drive page							
x+23								

GROUP

The Group drives (GROUP) bit is one, when it's desired to group the listed drives. This bit is set to zero when it's desired to ungroup the listed drives.

BAD

The Bad group (BAD) bit is only used with the MODE SENSE(10) command. This bit is undefined for the MODE SELECT(10) command, and its value is not checked.

When this bit is one this group of drives has a problem, and the group can not be accessed until the problem is corrected. Manual intervention may be required to correct the problem.

Number of drives

The Number of drives is used with the MODE SENSE(10) command to tell the RAID manager how many drives will be used to form this group. The minimum number of drives that can be grouped is two. The maximum number of drives that can be grouped is a value specified by the manufacturer. If zero drives, one drive or more than the maximum allowable number of drives are specified in this page, the RAID manager will return a CHECK CONDITION status and a sense key of TBD.

The RAID manager verifies that all of the drives of the group are of a suitable type and that all drives are READY. If the drives are not of a suitable type, the RAID manager will return a CHECK CONDITION status and a sense key of TBD. If any drive is NOT READY or the controller can not write to it the RAID manager will return a CHECK CONDITION status and the sense key will be whatever sense is returned from the drive., unless the drive is not selectable in which case the sense key will be TBD.

Stripe block size

For grouped drives that use drive striping, including RAID 5 drives, the stripe block size field defines the number of bytes that make up a stripe block. This field's unit is bytes.

The stripe block size is used to determine how the data will be written and read on the disks. The block sizes may be restricted to certain multiples of the individual drive's logical block size. A value of zero represents 64K.

Group properties

Bytes 8-11 of each Group Page specify the properties of the grouped drive using the following notation:

Table 4: Group properties

Bit Byte	7	6	5	4	3	2	1	0
8	Group property value							
9	Reserved							
11								

The group properties include:

Table 5: Group Property Values

Value	Description
TBDA	Striped Drive
TDBBn	Mirrored Drive of redundancy n (2-4)
TBD0	RAID 0
TBD1	RAID 1
TDB2	RAID 2
TBD3	RAID 3
TBD4	RAID 4
TBD5	RAID 5
All others	Reserved

Time Stamp and World Wide Names

The Time Stamp, Host World Wide Name, and Peripheral World Wide Name fields are written to the disks as part of the configuration information and are used to uniquely identify and verify the drives within a group.

The time stamp is a 32 bit integer whose value of time is in seconds since 00:00:00 GMT, January 1, 1970.

The Peripheral World Wide Name field contains the ID of the subsystem being addressed.

> NOTE: This field is filled in by the RAID manager when a MODE SELECT(10) command is executed and returned when a MODE SENSE(10) command is executed. The value in this field during a MODE SELECT(10) command is ignored.

After several bytes that are reserved for future use, a series of two or more Drive Pages appears. This Drive Page as the format that follows:

Table 6: Drive page format

Bit Byte	7	6	5	4	3	2	1	0
0	Entity address							
7								
8	Reserved							
9								
10	Device properties							
11	Reserved							
12	Drive serial number							
23								

Entity address

This field defines the entity address of the individual drive defined by this entry.

Drive properties

This field describes the properties of the drive and the functions it provides within the grouped drive according to the following table.

Table 7: Drive property values

Bit 7	Bit 6	Bit 5	Bit 4	Description	Bits 3-0	Value
0	0	0	0	Drive is group member	x	Reserved
1	x	x	0	Drive is standby	n	Order of use
x	1	x	0	Drive is mirror	n	Order of use
x	x	1	0	Drive is parity	n	Order of use

Drive Serial Number

This field contains the drive's serial number that is 12 bytes left justified ASCII.

The RAID shall return CHECK CONDITION status and shall set the sense key to ILLEGAL REQUEST if a reserved field contains any value other than zero.

If any of the drives identified to be in this drive group fails during the execution of the MODE SELECT(10) command with this page, then RAID shall return CHECK CONDITION status and set the sense key to TBD.

3.2 Internal operation mode page

Table 8: Internal operation mode page

Bit Byte	7	6	5	4	3	2	1	0
0	Reserved		Page Code (TBD1h)					
1	Reserved							
2	Page Length							
3								
4	Reserved		TB	ARBB	Reserved		APS	AES
5	Reserved							
7								
8	HVAC						FC	LoBt
9	Reserved							
11								

This page only applies to any addressable grouped drive.

When given in a MODE SELECT(10) command, the Internal Operation Mode Page is always saved and the PS bit is always ignored.

AES

The accumulate error statistics (AES) bit, if set to one, tells the grouped device to accumulate a vendor unique collection of error information. To extract this information from the grouped device, the initiator must issue the LOG SENSE command. The default is one.

APS

The accumulate performance statistics (APS) bit, if set to one, tells the grouped device to accumulate a vendor unique collection of performance information. To extract this information from the grouped device, the initiator must issue the LOG SENSE command. The default is one.

ARBB

The automatic reallocation of bad block (ARBB) bit, if set to one, tells the grouped device to automatically replace any bad data block if one is encountered during any READ or a WRITE command. If set to zero, the grouped device will report a status of CHECK CONDITION for any READ or WRITE command that produces a bad data block. The default is zero.

TBA

A transfer block (TB) bit of one indicates that a data block that is not recovered within the recovery limits specified shall be transferred to the initiator before CHECK CONDITION status is returned. A TB bit of zero indicates that such a data block shall not be transferred to the initiator. Data blocks that can be recovered within the recovery limits are always transferred, regardless of the value of the bit. This is the default value.

HVAC

The HVAC field is for heat, ventilation and air condition. It reflects the operation of the fans and whether the internal cooling of the enclosure is satisfactory. If sensors inside the RAID manager enclosure tell the manager that the temperature has exceeded a safe operational standard, than the manager will spin down the drives and disallow any access to the drives. This state remains in effect until this page with the TBD bit set is sent to the manager or the system performs a new power on sequence.

LoBt

The LoBt (Low Battery) bit is set to one if the manager detects that the batteries are low. When this event occurs, RAID manager turns off all capabilities that require normal battery state for proper operation and protection of data. This is not a settable parameter for the MODE SELECT(10) command.

FC

The FC (Fan Check) bit is set to one if the RAID manager can detect that the fans internal to the machine have stopped operating in its appropriate manner. Human intervention will be required to fix this situation. This is not a settable parameter for the MODE SELECT(10) command.

3.3 Drive state and information page

Table 9: Drive state and information page

Bit Byte	7	6	5	4	3	2	1	0
0	Reserved		Page code (TBD2h)					
1	Reserved							
2	Page length							
3								
4	Reserved							
7								
8	First drive page							
31								
	· ·							
n-23	Last drive page							
n								

The drive state and information page informs the initiator of the current status of all drives known to the addressed RAID manager at the time of the issuance of the command. This page does not contain any changeable parameters, so if it is requested in a MODE SELECT(10) command, a CHECK CONDITION status and sense key of ILLEGAL REQUEST is sent back to the initiator. This page must be specified in a MODE SENSE(10) command that is addressed to the RAID manager only. Otherwise the manager will return a CHECK CONDITION status and sense key of ILLEGAL REQUEST.

Table 10: Drive page format

Bit Byte	7	6	5	4	3	2	1	0
0	Reserved			SPN-DWN	CNR	DNR	NDS	NDF
1	Reserved						PCCW	CWE
2	GROUPED	MOVED	CONFLCT	Reserved				MRRB
3	Reserved							
4	Single drive entity address							
11								
12	Assigned group drive entity address							
19								
20	Reserved							
23								

NDF

The No Drive Found (NDF) bit contains a value of one if no drive was found at the location corresponding to the specified entity address. If the field contains a zero, then a drive does exist in this slot.

NDS

The No Drive Select (NDS) bit contains a value of one if the drive in this slot did not respond to selection. If the field contains a zero, then this drive is selectable.

DNR

The drive not ready (DNR) bit contains a value of one if the drive in this slot is not in the ready state. The RAID manager was able to select the drive and send an INQUIRY command, but a TEST UNIT READY command failed on this drive. If the field contains a zero, then this drive is ready.

CNR

The could not read (CNR) bit contains a value of one if the RAID manager could not read the configuration information from the drive in this slot. If the field contains a zero, then the controller is able to read data from this drive. During the RAID initialization the RAID controller attempts to read the configuration information off of each drive in order to determine which drives are part of a group.

SPN-DWN

The spun-down bit contains a value of one if the disk drive has been spun-down by the host using STOP UNIT SCSI command. This bit is zero otherwise. The STOP UNIT command may be required for diagnostic and maintenance activities.

PCCW

The per command cached write bit contains a value of one if the drive is managing cached writes on a per command basis. A bit in the FCP header will indicate whether to do a cached write or not. If the bit is zero, then the drive may be doing cached writes always or no cached writes at all.

CWE

The cached write enable bit contains a value of one if the drive is always doing cached writes. If the bit is zero, then the drive may be doing cached writes on a per command basis or no cached writes at all.

If both PCCW and CWE are set to zero, then cached writes are not allowed at all. See MODE SELECT Page TBD3 to see how these bits are set or cleared.

GROUPED

The Grouped drive (GROUPED) bit contains a value of one if this drive is a member of a grouped set of drives. If the Grouped field contains a value of zero, this drive is an individual drive.

MOVED

The Moved (MOVED) bit is only defined for grouped drives. The Moved (MOVED) bit is a one if the current location of the drive is different than the location of the drive as saved in the configuration information. This bit is zero otherwise.

CONFLCT

The Conflict (CONFLCT) bit is only defined for grouped drives. The Conflict bit is set to one if two or more groups claim the drive as a member of their group.

MRRB

The most recent reallocated block (MRRB) bit is set to one if this individual drive was affected by the last REASSIGN BLOCK command that was executed while using this group. This bit is an aid to any initiator that has sent a REASSIGN BLOCK command to the grouped drive and wants to see which drive within the set actually did the reallocation of the block.

Single Drive Entity Address

The Single Drive Entity Address defines the actual entity address of the single drive and can be used to physically locate the drive in a subsystem in a vendor unique manner.

Assigned Group Drive Entity Address

The assigned Group Address is the address that is used to access the group of which the drive is a member, for example when reading or writing. The address of a group is assigned by the controller when the group is formed.

3.4 Cached write configuration page

Table 11: Cached write configuration page

Bit Byte	7	6	5	4	3	2	1	0
0	Reserved		Page code (TBD3h)					
1	Reserved							
2	Page length (0Eh)							
3								
4	GROUP	ALL	Reserved					
5	Reserved							
6	Assigned group drive entity address							
13								
14	PURGE	Reserved					PCCW	CWE
15	Reserved							
17								

This page only applies to the RAID manager.

When given in a MODE SELECT(10) command, the Cached Write Configuration Page is always saved and the PS bit is always ignored.

The Cached Write Configuration page defines the parameters that affect how WRITE commands function in the grouped device. The term 'Cached Write' means that grouped drive will return GOOD status for a WRITE command after successfully receiving the data and prior to having successfully written it to the medium.

This page is writable via the MODE SELECT(10) command when addressed to a RAID manager, a grouped drive, or an individual drive. This page is readable via the MODE SENSE(10) command when addressed to a RAID manager, a grouped drive or an individual drive. The parameters savable (PS) bit is only used with the MODE SENSE(10) command. This bit is reserved with the MODE SELECT(10) command. A PS bit of one indicates that RAID is capable of saving the page in a nonvolatile vendor-specific location.

GROUP

A RAID drive is selected by setting the port and target address of the drive in the Port Address and Target Address fields respectively. If the RAID drive is a group drive, then the GROUP bit is set to one. If the GROUP bit is set to zero, then the port and target address is assumed to be an individual drive.

ALL

The ALL bit tells the RAID to enable cached writes for all drives managed by the addressed RAID manager. All other bits are ignored if ALL is set to one. If ALL is set to one in a previous MODE

SELECT(10) command, then a second MODE SELECT(10) command has the ALL bit set to zero, then the cached write mode is disabled for all devices.

The purge (PURGE) bit, cached write enable (CWE) bit, and per command cached write (PCCW) bit form a three bit field that has the following definition:

Table 12: Cache control bit definitions

PURGE	PCCW	CWE	Description
0	0	0	No cached writes allowed to this drive
0	0	1	Do cached writes to this drive
0	1	x	Follow the command request bit for this drive
1	x	x	Purge any outstanding write data

PCCW = 0, CWE = 0

If PCCW is zero and CWE is zero, then no cached writes will be allowed to this drive even if the TBD bit in the fibre channel packet states otherwise, and responses for every WRITE command for this drive won't be sent to the initiator until the data is on the media.

PCCW = 0, CWE = 1

If PCCW is zero and CWE is one, then the initiator allows every WRITE command to this drive to be a cached write even if the TBD bit in the fibre channel packet states otherwise.

PCCW = 1, CWE = DON'T CARE

If PCCW is set to one, then the RAID will always check the TBD bit in the fibre channel packet for each WRITE command to this drive to decide if the write should be a cached write or not.

PURGE

On the occasion that the drive has left the enclosure and data for the drive exists in the RAID and the RAID has told the host that this event has occurred, some host may want to throw away the data. This will occur if the PURGE bit is set to one in this page. Setting the PURGE bit to one must be done under extreme caution as when it is set, the data that is tossed will be gone forever.

If PURGE is set to zero, the RAID will not consider this to be an error.