SCSI-3 Storage Array Tutorial

IBM Storage Subsystems Products
Rochester, Minnesota
George Penokie

Internet ID:
GOP@RCHVMP3.VNET.IBM.COM

SCSI-3 Controller Command Standard What is it?

- A model describing objects and the relationships between those objects within storage arrays
  - Sections one through five and the annexes of the SCC Standard

- Command set
  - Uses SCSI structures (CDBs, Parameters list, etc.)
  - Section six of the SCC Standard

- Command set is dependent on the model but the model does not require a SCSI command set

Terminology

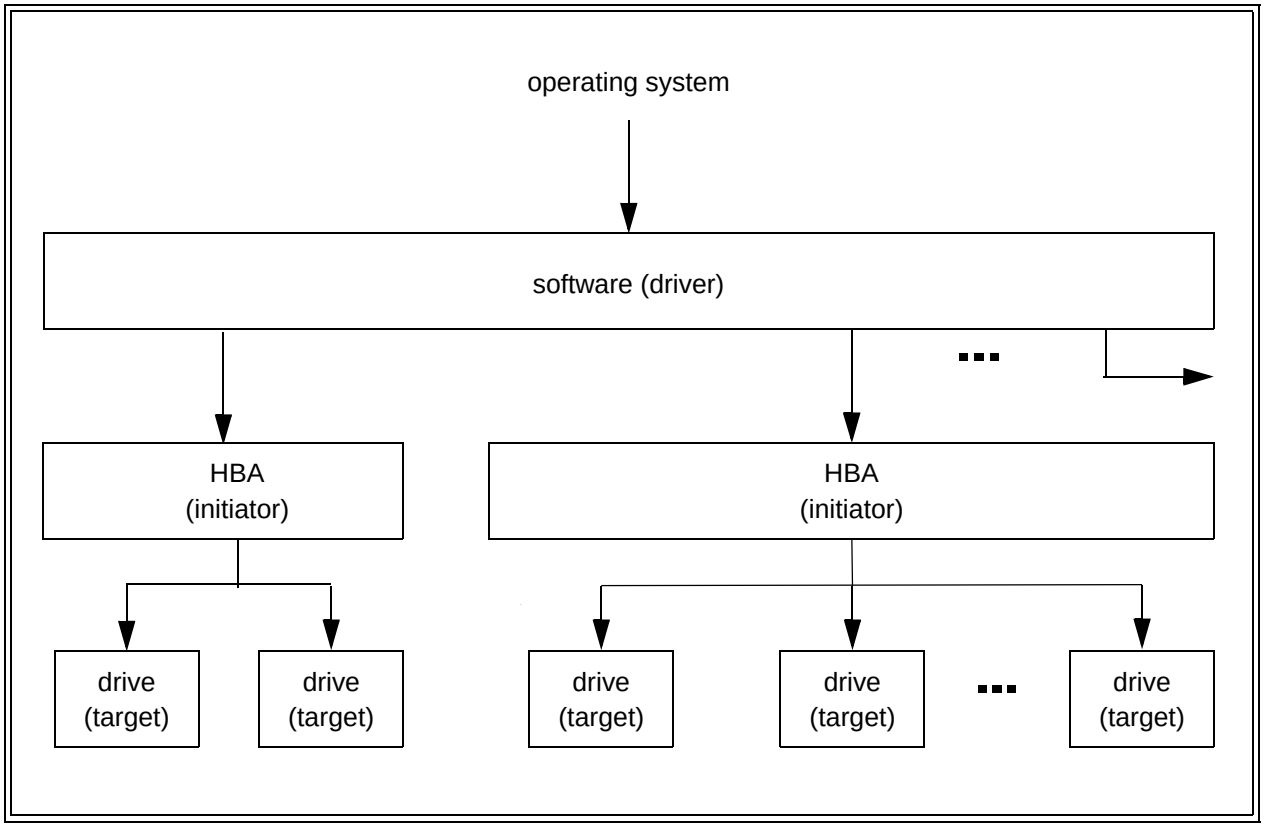
- SCSI-3 Storage Array
  - A SCSI-3 device that controls multiple SCSI devices using the rules and commands defined in the SCSI-3 Controller Commands Standard
  - Contains one or more SACLs

- Storage Array Conversion Layer (SACL)
  - Translates input logical unit numbers into one or more output logical unit numbers
  - Routes logical block addresses to one or more logical units
  - May convert input logical block addresses to output logical block addresses

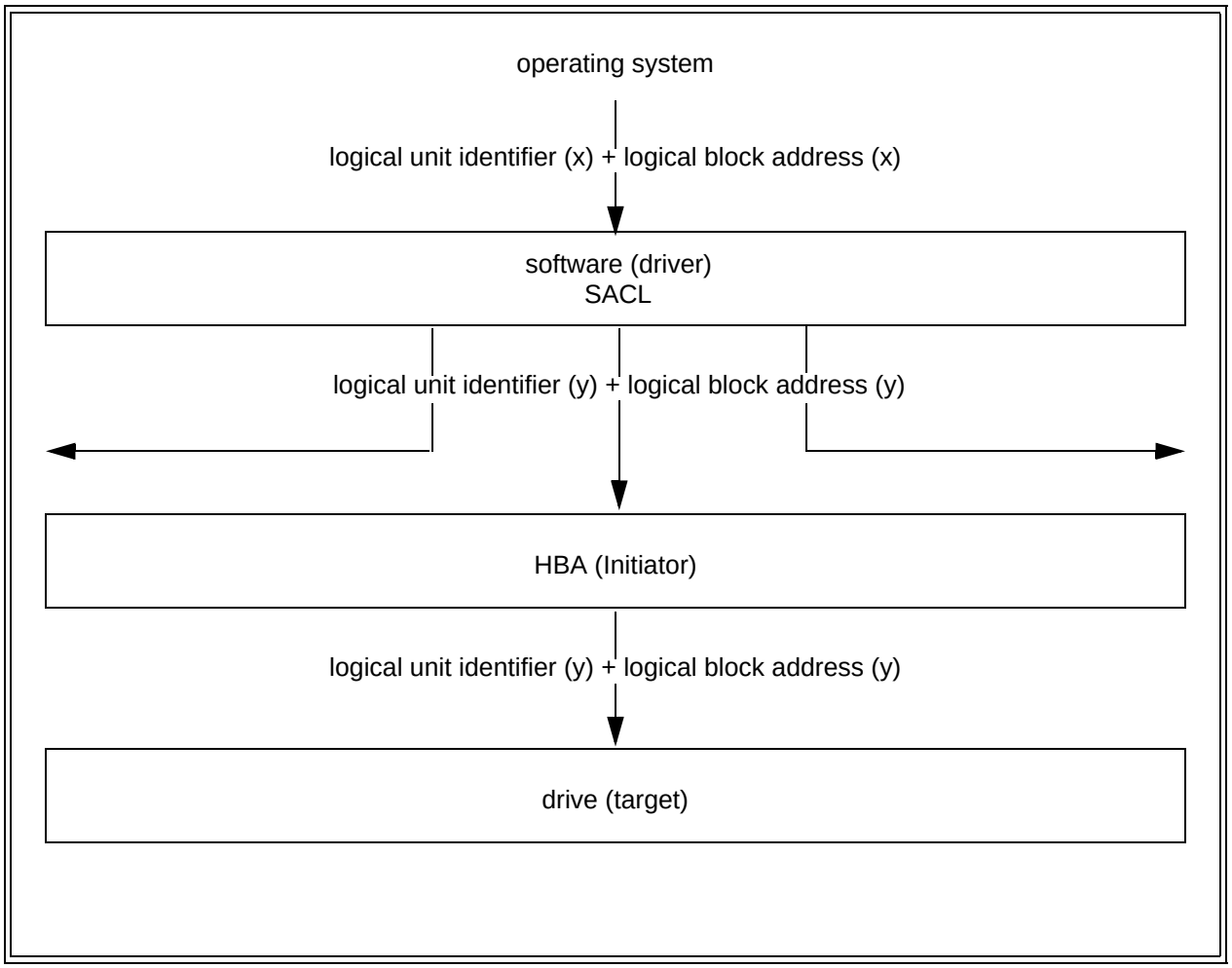
System Layering

- A system may contain zero or more SCSI Storage Arrays
- A system may contain one or more SACLs
  - May exist anywhere within the system

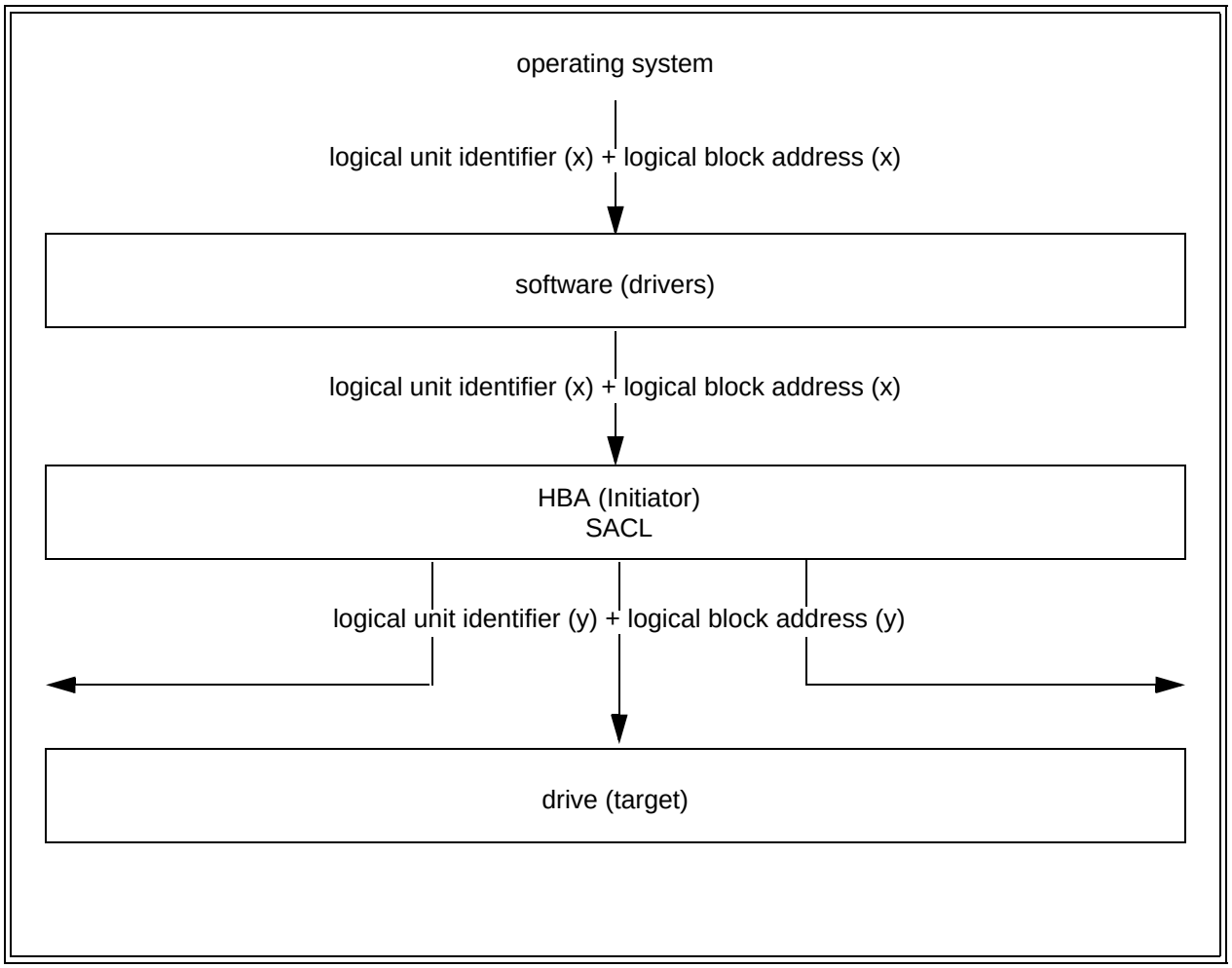
Typical System



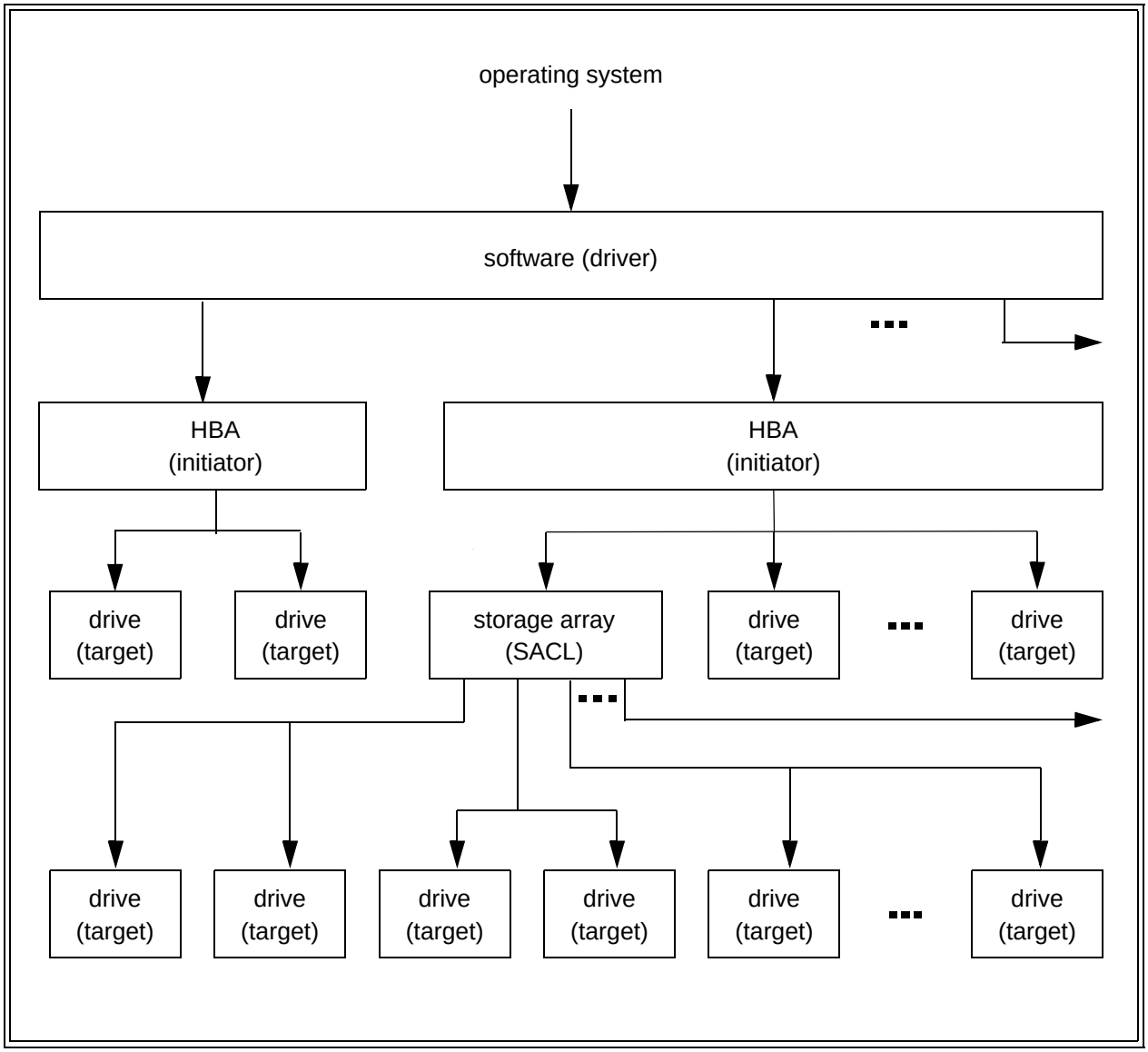
Software SACL Branch



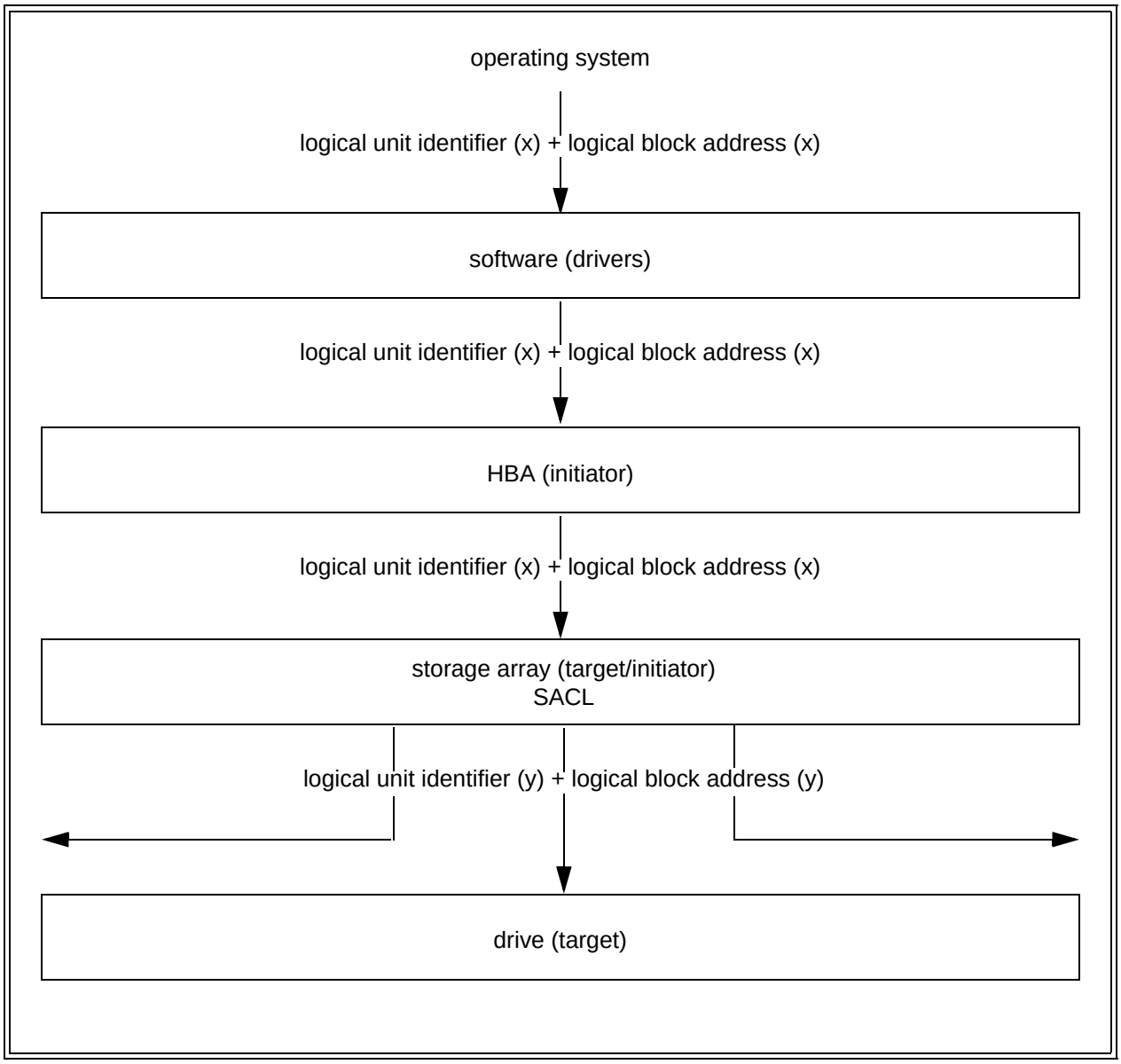
Host Bus Adapter (HBA) SACL Branch



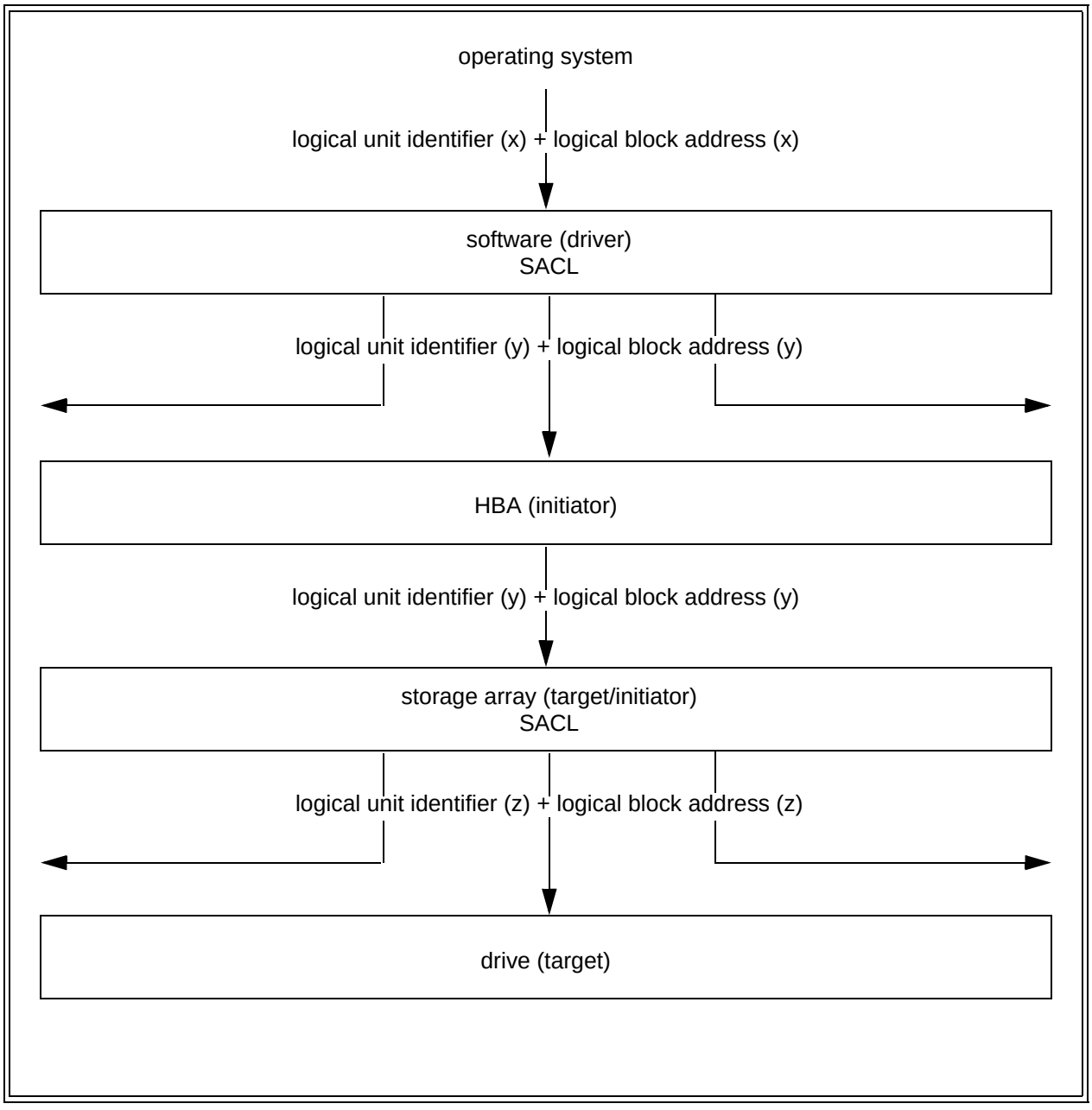
System With A SCSI-3 Storage Array Attached



SCSI-3 Storage Controller SACL Branch



Multiple SACLs within one Branch



SCSI-3 Storage Array Physical Objects

- Peripheral device: Any device identifiable as a SCSI peripheral device type:
  - Disk drives
  - Tape drives
  - Optical devices
  - etc.

- Component device: Any addressable device not identifiable as a SCSI peripheral device:
  - Controller electronics that contain a SACL
  - Non-volatile cache
  - Power Supply
  - Display
  - Keypad entry
  - Fan

SCSI-3 Storage Array Logical Objects

- P_extent: A contiguous block of logical block addresses on a single peripheral device.
  - A single p_extent cannot exist on more than one peripheral device
  - Used to:
    - Create redundancy groups
    - Modify redundancy groups
    - Create spares

- Ps_extent: A contiguous block of logical block addresses on a single peripheral device that excludes any logical blocks addresses identified as check data.
  - A single ps_extent cannot exist on more than one peripheral device
  - Ps_extents are a result of the creation of redundancy groups
  - Used to:
    - Create volume sets
    - Modify volume sets

SCSI-3 Storage Array Logical Objects (Cont.)

- Redundancy group: A grouping of one or more p_extent(s) that have a common type of protection
  - The logical block addresses of two or more redundancy group may overlap if the overlapping logical block addresses do not contain any check data

- Volume set: A grouping of one or more ps_extent(s) that provide a contiguous range of logical block addresses for reading and writing user data
  - The logical block addresses of volumes sets cannot overlap

- Spare: A p_extent, peripheral device, or component that will be automatically exchanged with a like object if that object fails

Addressing the SCSI-3 Storage Array

- Objects are directly or indirectly addressable
  - Directly addressable objects are:
    - Peripheral devices
    - Volume sets
  - Indirectly addressable objects are:
    - Peripheral devices
    - Volume sets
    - Component devices
    - Redundancy groups
    - Spares

Direct Addressing

- Any command may be sent
- Uses an 8-byte field split into 4 four 2-byte fields to address up to four levels of objects
- Each 2-byte field identifies and locates objects within a level
- The 2-byte field contains the address method to be used
  - Peripheral device addressing method
    - 63 buses
    - 256 peripheral devices per bus
  - Volume set addressing method
    - $2^{14} - 1$ volume sets

Format of 2-byte field

- Peripheral device addressing format:

Bit / Byte	7	6	5	4	3	2	1	0
n-1	0	0	BUS NUMBER					
n	TARGET/LUN							

- Volume set addressing format:

Bit / Byte	7	6	5	4	3	2	1	0
n-1	0	1	(MSB)					
n	LUN							(LSB)

Format of 8-byte field

■

Bit Byte	7	6	5	4	3	2	1	0
0	FIRST LEVEL ADDRESSING							
1								
2	SECOND LEVEL ADDRESSING							
3								
4	THIRD LEVEL ADDRESSING							
5								
6	FOURTH LEVEL ADDRESSING							
7								

The Missing Bus

- A bus number of zero in the 2-byte field for peripheral device addressing addresses the SCSI-3 storage array directly

SCSI-3 Storage Array Base Address

- All SCSI-3 storage arrays have a base address
- The base address is logical unit number zero
- A value of 0000h in the 2-byte field for peripheral device addressing will address the base address
- All commands to objects that require indirect addressing are sent to the base address

Addressing Exception for SIP

- The Identify message and a mode page replace the 8-byte and 2-byte fields

- If there is no active LUN mapping mode page the Identify message contains the address method to be used
  - Peripheral device addressing method
    - VOLSEL field set to zero
    - 32 peripheral devices
  - Volume set addressing method
    - VOLSEL field set to one
    - 32 volume sets

- If there is an active LUN mapping mode page the Identify message contains
  - Base address addressing
    - Bits 5-0 set to zero
  - A pointer to one of 31 8-byte addressing fields
    - VOLSEL field set to zero
  - Volume set addressing method
    - VOLSEL field set to one
    - 32 volume sets

Format of the Identify Message and LUN Mapping Mode Page

- Format of Identify Message

Bit / Byte	7	6	5	4	3	2	1	0
0	IDENTIFY	DISCPRIV	VOLSEL	LUN				

- Format of LUN mapping mode page

Bit / Byte	7	6	5	4	3	2	1	0
0	PS	RESERVED	PAGE CODE (xxh)					
1	PAGE LENGTH (FAh)							
2	RESERVED							
3	RESERVED							ACTIVE
4	(MSB)	LUN 1 MAPPING						
11								(LSB)
	. . .							
244	(MSB)	LUN 31 MAPPING						
251								(LSB)

Notation for Addressing Examples

The conventions used within the examples are:

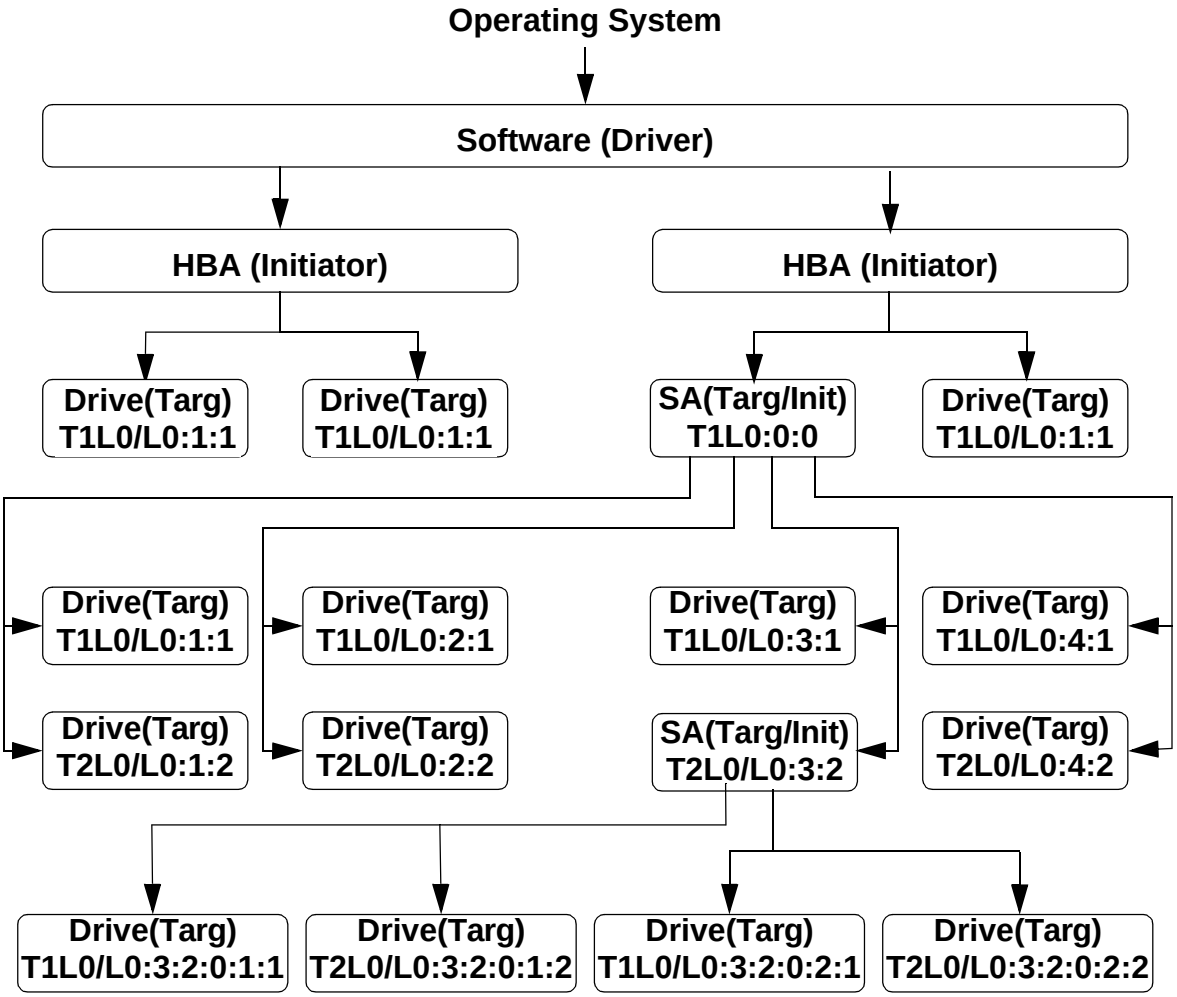
Layer 1 M:P:T or M:L or u
Layer 2 M:P:T or M:L or u
Layer 3 M:P:T or M:L or u
Layer 4 M:P:T or M:L or u

Where:

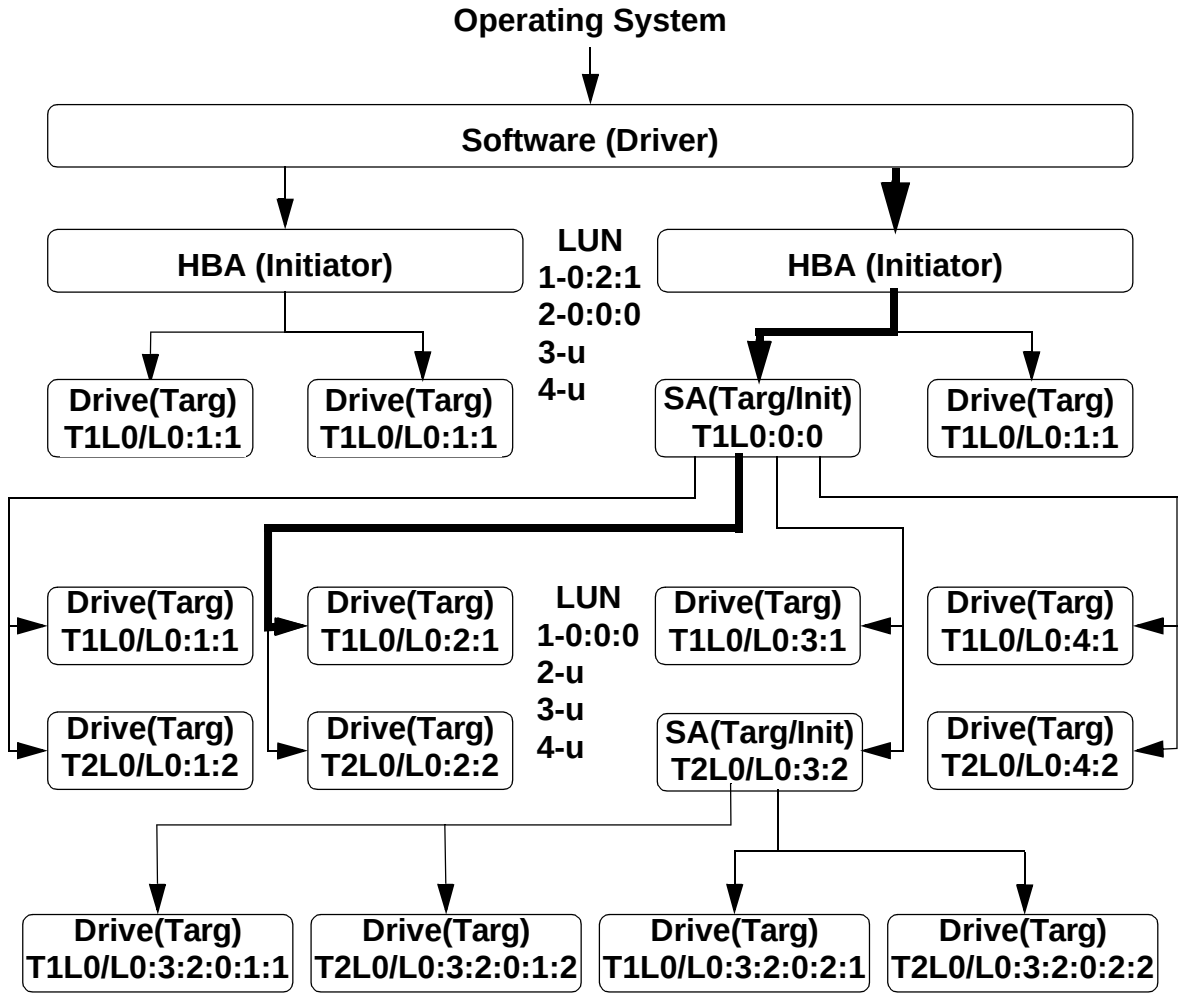
M is the Address Method (2 bit field)
P is the Bus Number (6 bit field)
T is the Target (8 bit field)
L is the Logical Unit Number (14 bit field)
u means unused and set to zero (16 bit field)

Note: All of the examples use the peripheral device addressing method therefore M=0

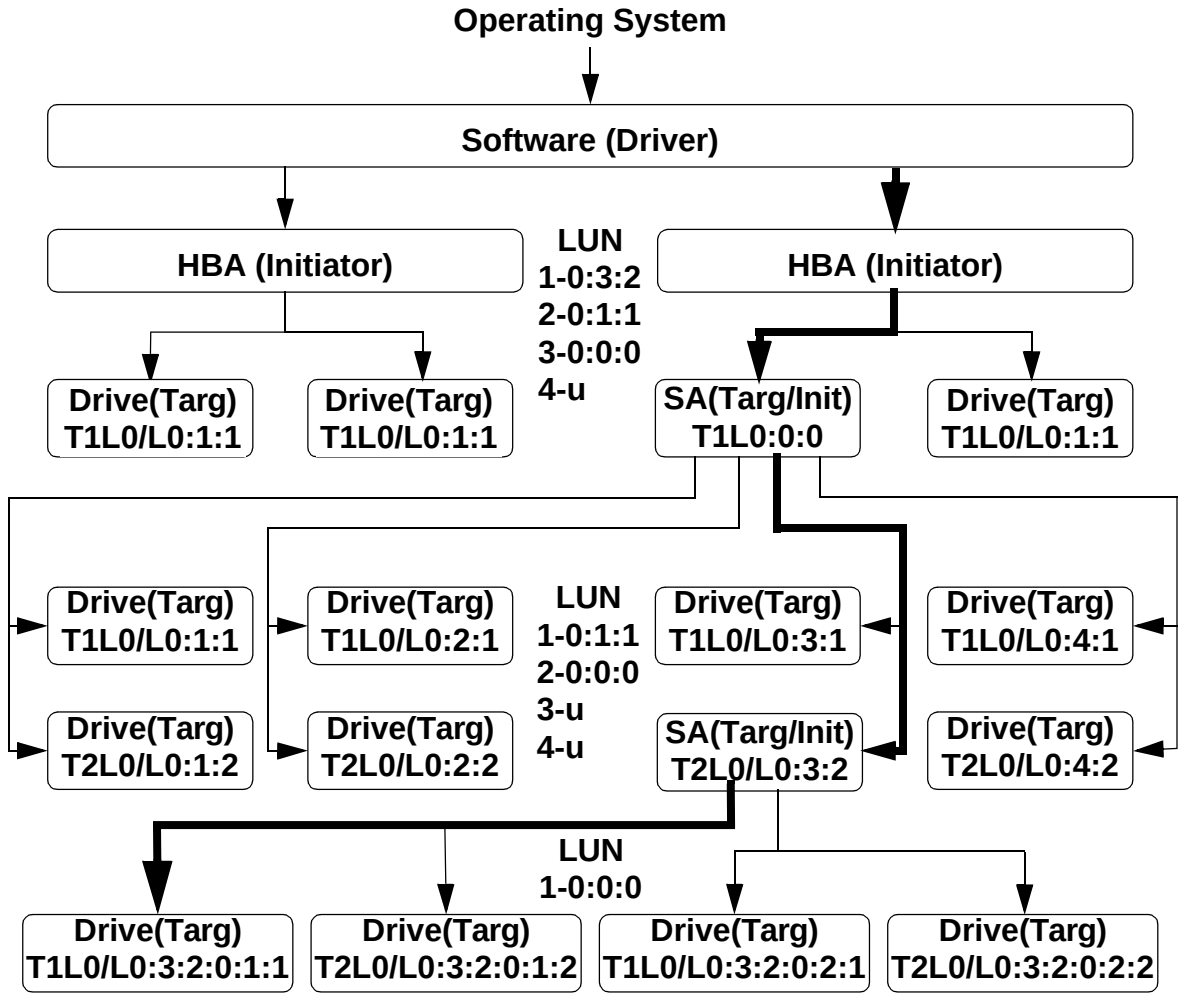
Addressing


Operating System
Software (Driver)
HBA (Initiator)
HBA (Initiator)
Drive(Targ) T1L0/L0:1:1
Drive(Targ) T1L0/L0:1:1
SA(Targ/Init) T1L0:0:0
Drive(Targ) T1L0/L0:1:1
Drive(Targ) T1L0/L0:1:1
Drive(Targ) T1L0/L0:2:1
Drive(Targ) T1L0/L0:3:1
Drive(Targ) T1L0/L0:4:1
Drive(Targ) T2L0/L0:1:2
Drive(Targ) T2L0/L0:2:2
SA(Targ/Init) T2L0/L0:3:2
Drive(Targ) T2L0/L0:4:2
Drive(Targ) T1L0/L0:3:2:0:1:1
Drive(Targ) T2L0/L0:3:2:0:1:2
Drive(Targ) T1L0/L0:3:2:0:2:1
Drive(Targ) T2L0/L0:3:2:0:2:2

Example 1:Address Drive at Level 1

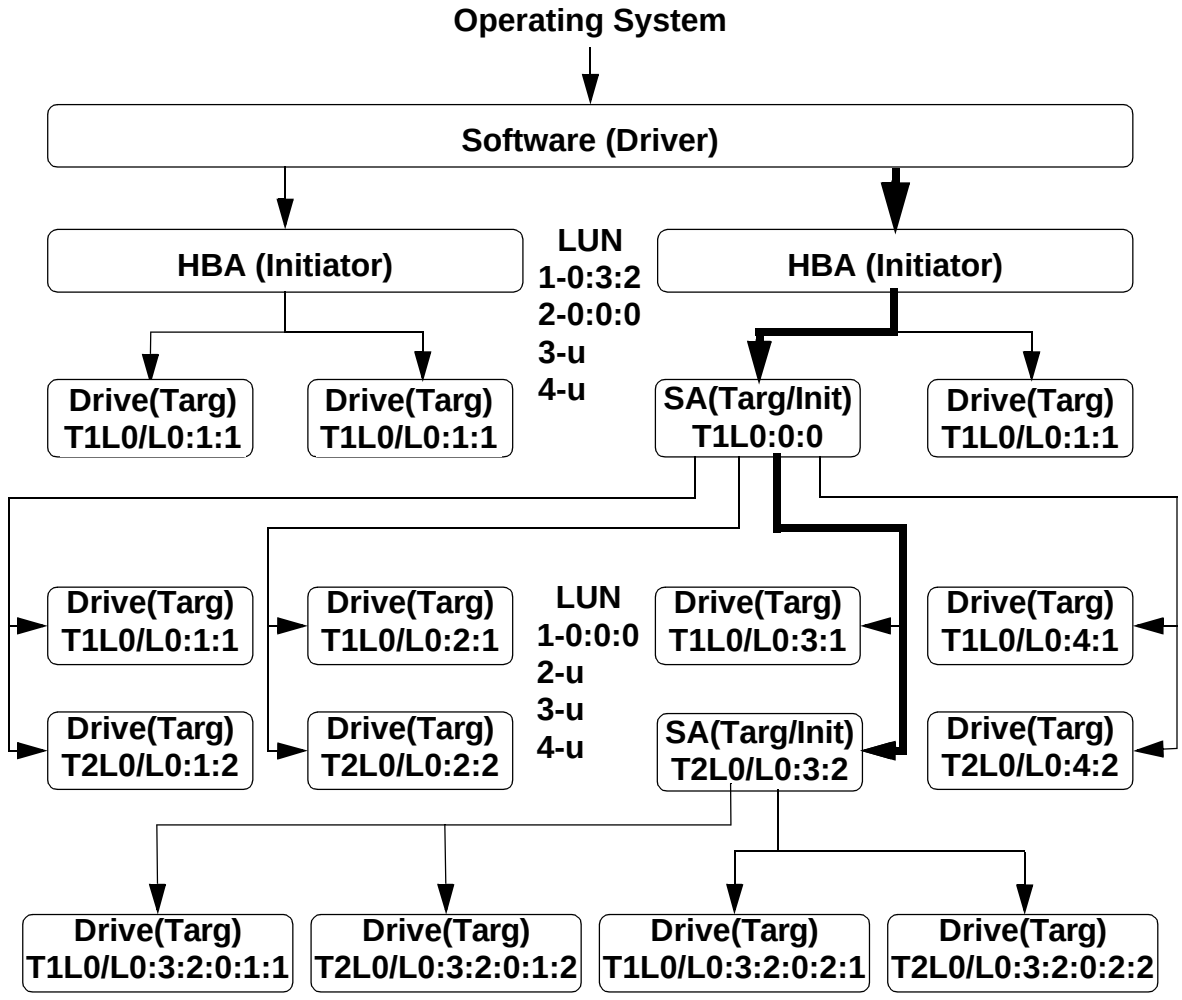


Example 2:Address Drive at Level 2



Example 3:Address Fan at Level 1

■



Indirect Addressing

- All indirect addressing fields are contained within commands sent to a SCSI-3 storage arrays' base address
- Only commands defined by the SCC standard use indirect addressing
- Any commands that contain indirect addressing must be sent to the base address of a SCSI-3 storage array
- A Logical Unit Type (4-bits) field contains type of object to address
  - Peripheral devices
  - Volume sets
  - Component devices
  - Redundancy groups
  - Spares
- A LUN (2-byte) field contains the logical unit number of the object to address

Format on Indirect Addressing Fields

■ Data format of LOGICAL UNIT DESCRIPTOR

Bit / Byte	7	6	5	4	3	2	1	0
0	RESERVED							
1	RESERVED				LOGICAL UNIT TYPE			
2	(MSB) LUN							
3								(LSB)

Proprieties of Objects

- Protected objects
  - An object that can tolerate one or more objects failing without any loss of user data or availability
  - Objects that can be protected by using spares
    - Component devices
    - Peripheral device
    - P_extents
  - Objects that can be protected by using redundancy groups
    - Volume sets

- Association of objects
  - A linking of like objects to form an object
    - P_extents become associated with a redundancy group during the creation or modification of that redundancy group
    - Redundancy groups become associated with a volume set during the creation or modification of that volume set

Proprieties of Objects (Cont.)

- Attachment of objects
  - The linking of objects to component devices
  - Any of the following objects may be attached to a component device:
    - Peripheral devices
    - Volume sets
    - Component devices
    - Redundancy groups
    - Spares
  - The behavior of attachments and their interactions with component devices are not defined in the SCC standard

- Covering of objects
  - The protection of objects using spares
  - Only like objects can be covered
  - Any of the following objects may be covered by spares:
    - Peripheral devices
    - P_extents
    - Component devices
  - When an object is covered it assumes all the characteristics of the failed object

Operations on Objects

- Adding objects
  - Adding an object makes it addressable
  - The following object may be added:
    - Peripheral devices
    - Component devices

- Exchanging objects
  - Replacing an object with a like object
  - Only like objects can be exchanged
  - Any of the following objects may be exchanged:
    - Peripheral devices
    - P_extents
    - Component devices
  - When an object is exchanged it assumes all the characteristics of the object it is replacing

Operations on Objects (Cont.)

- Removing objects
  - Removing objects makes them no longer addressable:
  - Any of the following objects may be removed:
    - Peripheral devices
    - Component devices
    - Redundancy groups
    - Volume sets
    - Spares
  - Any logical block addresses within a removed volume set become unassigned protected space
  - Any logical block addresses within removed redundancy group or spare become unassigned p_extents
  - The removing of a component spare removes the covering, however, the component remains addressable

- Rebuilding objects
  - The rebuild operation recreates protected space contents or any check data within a p_extent using check data and protected space contents from the remaining p_extents within the redundancy group
  - The regenerated protected space contents or any recalculated check data shall be written to the p_extent being rebuilt

Operations on Objects (Cont.)

- Recalculating objects
  - The recalculate operation recreates check data from protected space contents
  - The recreated check data shall be written to the check data location being recalculated

- Regenerating objects
  - The regenerate operation recreates inaccessible protected space contents from accessible check data and protected space contents
  - The recreated protected space contents is not saved

- Verifying objects
  - The verify operation recreates check data from protected space contents and compare the recreated check data with the current check data
  - If the recreated check data does not match the current check data an exception condition shall be created

SCSI-3 Storage Array States

- Gives current operating condition of selected logical unit

- Base address states
  - Readying state: Indicates if any logical units within the SCSI-3 storage array being initialized and access is limited.
  - Non-addressable component failure state: Indicates one or more non-addressable part(s) have failed. (e.g. power supply failure, LED failure, cache failure, etc. that are not defined as component devices).
  - Abnormal state: Indicates one or more addressable devices within the SCSI-3 storage array are indicating a state other than available.

SCSI-3 Storage Array States (Cont.)

- Volume set states

Codes	States	Description
00h	Available	The addressed volume set is operational.
01h	Broken	The addressed volume set is capable of being supported but it has failed.
02h	Data lost	Within the addressed volume set data has been lost.
03h	Exposed	Within the addressed volume set data is not protected. In this state all data is still valid, however, a failure causes a loss of data or a loss of data availability.
0Ch	Protection disabled	Within the addressed volume set the generation of check data has been disabled. In this state all data is still valid, however, a failure causes a loss of data or a loss of data availability.
04h	Partially exposed	Within the addressed volume set one or more logical unit(s) have failed. In this state all data is still protected.
05h	Protected rebuild	One or more of the redundancy groups associated with the addressed volume set is in the process of a rebuild operation. In this state all data is protected.
06h	Not available	The addressed volume set is capable of being supported but has not been configured.
07h	Not supported	The addressed volume set is not capable of being configured.
08h	Readying	The addressed volume set is being initialized and access to the volume set is limited.
09h	Rebuild	One or more of the underlaying redundancy groups associated with the addressed volume set is in the process of a rebuild operation. In this state data is not protected.
0Ah	Recalculate	The addressed volume set is in the process of a recalculate operation.
0Bh	Spare in use	Within the addressed volume set a spare is being used. In this state all data is still protected.
0Dh	Verify in progress	Within the addressed volume set data is being verified.
0Eh-3Fh	Reserved	
40h-7Fh	Vendor Specific	

SCSI-3 Storage Array States (Cont.)

- Redundancy group states

Codes	States	Description
00h	Available	The addressed redundancy group is configured.
01h	Exposed	Within the addressed redundancy group data is not protected. In this state all data is still valid, however, a failure causes a loss of data or a loss of data availability.
02h	Invalidated Protected Space	Within the addressed redundancy group data has been lost. In this state the protected space is no longer intact.
03h	Not Available	The addressed redundancy group is capable of being supported but has not been configured.
04h	Not Supported	The addressed redundancy group is not capable of being configured.
05h	Partially Exposed	Within the addressed redundancy group one or more logical unit(s) have failed. In this state the protected space is protected.
0Ah	Protection disabled	Within the addressed redundancy group the generation of check data has been disabled. In this state all data is still valid, however, a failure causes a loss of data or a loss of data availability.
06h	Present	The addressed redundancy group is present but no other status is available.
07h	Protected Rebuild	The addressed redundancy group is in the process of a rebuild operation. In this state the protected space is protected.
08h	Rebuild	The addressed redundancy group is in the process of a rebuild operation. In this state the protected space is not protected.
09h	Recalculate	The addressed redundancy group is in the process of a recalculate operation.
0Bh	Verify in progress	Within the addressed redundancy group data is being verified.
0Ch-3Fh	Reserved	
40h-7Fh	Vendor Specific	

SCSI-3 Storage Array States (Cont.)

- Peripheral device and p_extent states

Codes	States	Description
00h	Available	The addressed peripheral device or p_extent is operational.
01h	Broken	The addressed peripheral device or p_extent is capable of being supported but it has failed.
02h	Not available	The addressed peripheral device or p_extent is capable of being supported but no device is connected.
03h	Not supported	The target is not capable of supporting a device at the addressed peripheral device or p_extent.
04h	Present	The addressed peripheral device or p_extent is present but no other status is available.
05h	Readying	The addressed peripheral device or p_extent is being initialized and access to the peripheral device or p_extent is limited.
06h-3Fh	Reserved	
40h-7Fh	Vendor Specific	

- Spare states

Codes	States	Description
00h	Available	The addressed spare is operational.
01h	Broken	The addressed spare is capable of being supported but it has failed.
02h	Not available	The addressed spare is capable of being supported but has not been configured.
03h	Not supported	The addressed spare is not capable of being configured.
04h	Present	The addressed spare is present but no other status is available.
05h	Spare in use	The addressed spare is being used.
06h-3Fh	Reserved	
40h-7Fh	Vendor Specific	

SCSI-3 Storage Array States (Cont.)

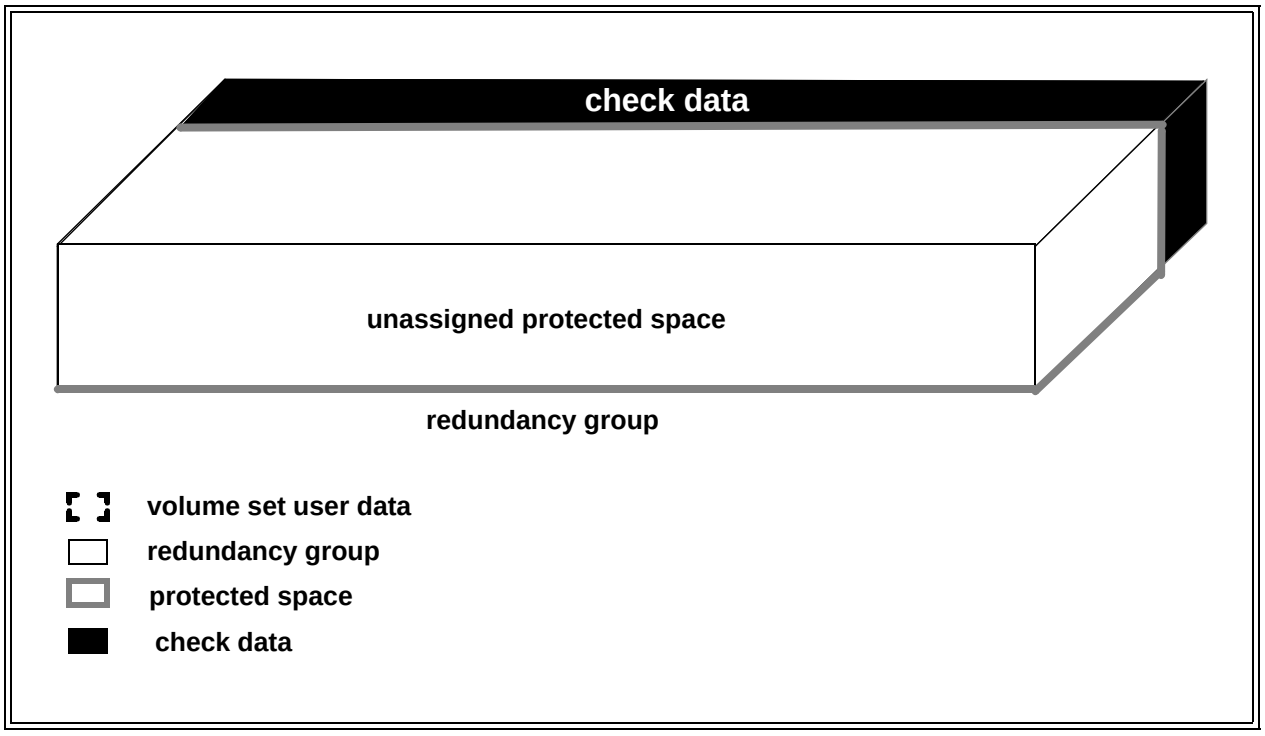
- Component states

Codes	States	Description
00h	Available	The addressed component device is fully operational.
01h	Broken	The addressed component device is capable of being supported but it has failed.
02h	Reserved	
03h	ITTU	The addressed component device is the reporting component device. This state shall not be reported unless the command allows the reporting of multiple states. More that one component device may report an ITTU state in a single state request.
04h	Not available	The addressed component device is capable of being supported but no component is present.
05h	Not supported	The target is not capable of supporting a component on the addressed component device.
06h	Present	The addressed component device is present but no other status is available.
07h	Readying	The addressed component device is being initialized and access to the component device is limited.
08h-3Fh	Reserved	
40h-7Fh	Vendor Specific	

Exception Conditions

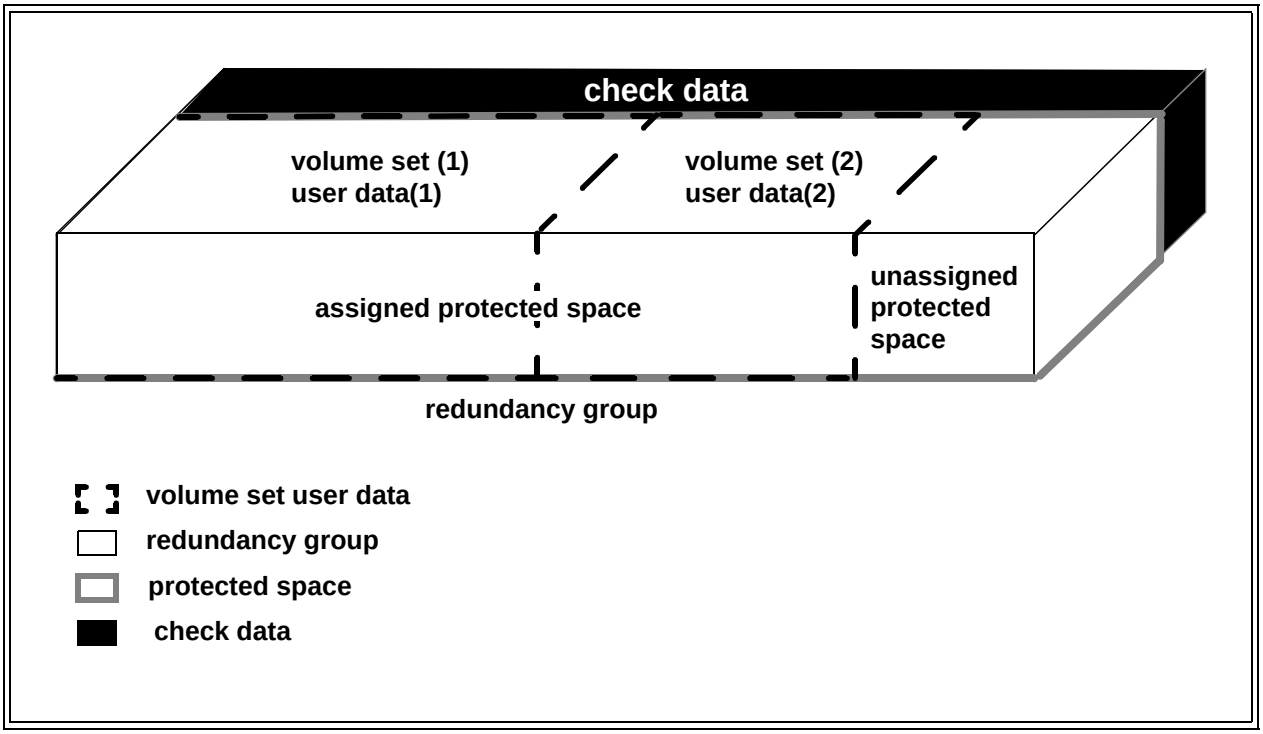
- Exception conditions indicate the following:
  - a change occurred in the physical configuration,
  - a change occurred in a volume set configuration,
  - a change occurred in a redundancy group configuration,
  - a change occurred in a spare,
  - a change occurred in the operation state of the SACL,
  - a repair action is requested (e.g. device is predicting failure),
  - a repair action is required to restore the volume sets availability (e.g. power supply failure),
  - a repair action is required to restore the volume sets level of integrity (e.g. device fails), or
  - an error occurred.

Single Redundancy Group Example



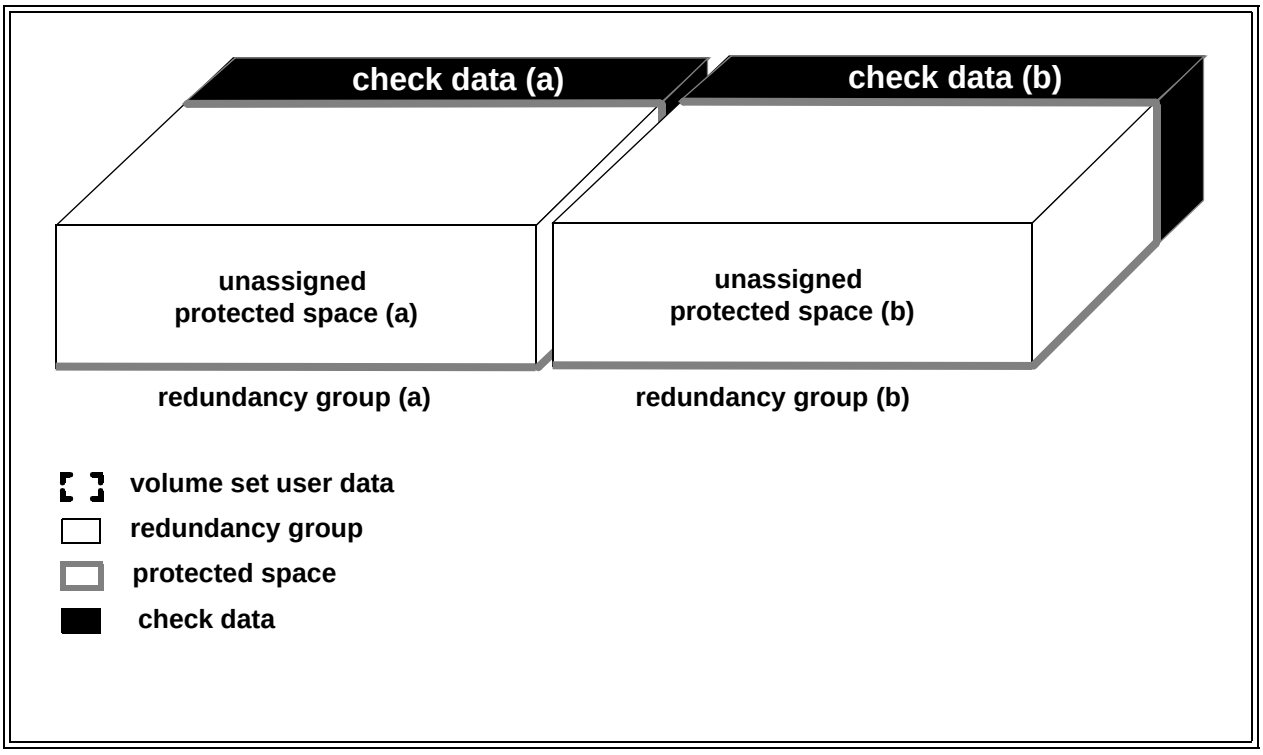
Multiple Volume Set Associated with a Single Redundancy Group Example

■



Multiple Redundancy Groups Example

■



Single Volume Set Associated with Multiple Redundancy Groups