

File Systems and Thin Provisioning

Frederick Knight
NetApp

Thin Provisioning

- ❑ What is Thin Provisioning?
- ❑ Why is it important?
- ❑ How does the host interact with SCSI TP?
- ❑ How does the host interact with ATA TP?

What is Thin Provisioning

- ❑ Space efficiency technique
 - ❑ Reduces physical capacity, power, cost, footprint
- ❑ Based on historical data usage models
 - ❑ File systems are never 100% full
 - ❑ Databases are never 100% full
- ❑ Allows storage device to report larger capacity than actually exists

Why is it important?

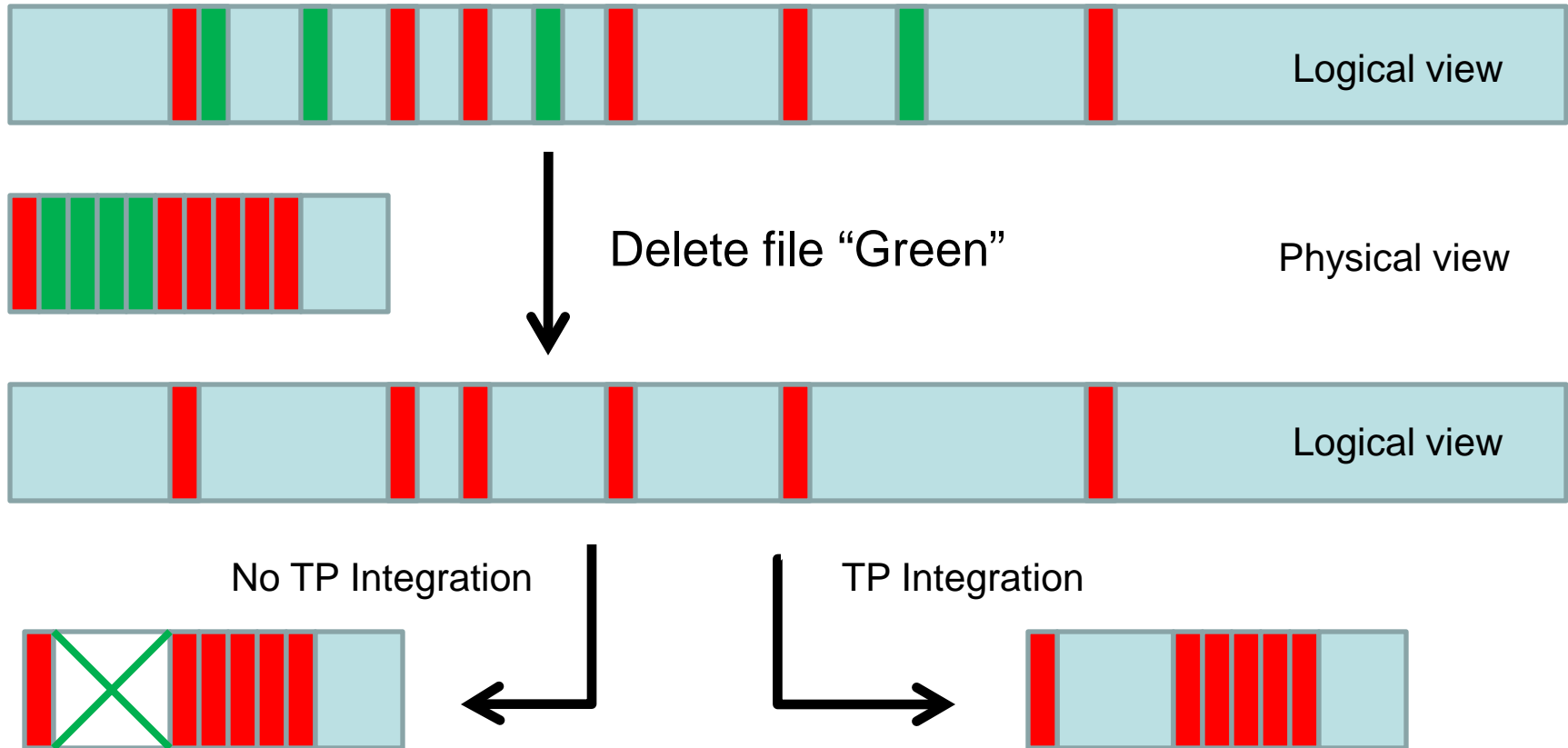
- ❑ Space efficiency technique
 - ❑ Reduces physical capacity
 - ❑ Reduces footprint
 - ❑ Reduces power
 - ❑ Reduces cost
- ❑ Buy only what you need
- ❑ Power, Cool, and Manage only what you need

How hosts interact with TP

- ❑ Defined in T10 SBC-3 standard
 - ❑ For SCSI Devices
 - ❑ Called “Logical Block Provisioning”
- ❑ Defined in T13 ACS-2 standard
 - ❑ For ATA and SATA Devices
 - ❑ Called “Data Set Management – TRIM”

- ❑ Host integration is important

Why hosts integration matters



SCSI Provisioning Model

- ❑ Full Provisioning
 - ❑ Static physical space always exists
- ❑ Resource Provisioning
 - ❑ Dynamic Logical to Physical remapping
- ❑ Thin Provisioning
 - ❑ May not have sufficient physical space
 - ❑ Transient / Persistent out of space conditions
 - ❑ Resource reporting

- ❑ UNMAP command
- ❑ WRITE SAME command
- ❑ GET LBA STATUS command
- ❑ Operational Parameters
 - ❑ VPD pages (static read only state)
 - ❑ Log pages (dynamic read only state)
 - ❑ Mode pages (dynamic read write state)

SCSI Commands - UNMAP

- ❑ No write data transferred
- ❑ Multiple Starting LBAs + Length (in parameter data)

Byte	Bit	7	6	5	4	3	2	1	0	
0	(MSB)	UNMAP DATA LENGTH (n-1)								(LSB)
1										
2	(MSB)	UNMAP BLOCK DESCRIPTOR DATA LENGTH (n-7)								(LSB)
3										
4		Reserved								
.....										
7										
UNMAP block descriptors										
0	(MSB)	UNMAP LOGICAL BLOCK ADDRESS								(LSB)
.....										
7										
8	(MSB)	NUMBER OF LOGICAL BLOCKS								(LSB)
.....										
11										
12		Reserved								
.....										
15										

Multiple Instances

SCSI Commands – WRITE SAME

- ❑ Must transfer one logical block of write data
- ❑ Single starting LBA + Length (in the CDB)

Byte	Bit	7	6	5	4	3	2	1	0
0		OPERATION CODE (93h)							
1		WRPROTECT		ANCHOR	UNMAP	PBDATA	LBDATA	Reserved	
2	(MSB)	LOGICAL BLOCK ADDRESS							
.....									
9		(LSB)							
10	(MSB)	NUMBER OF LOGICAL BLOCKS							
.....									
13		(LSB)							
14		Reserved			GROUP NUMBER				
15		CONTROL							

Only ONE

SCSI Commands – WRITE SAME

- ❑ WRITE SAME is a WRITE command
- ❑ READ must return the data written
- ❑ Additional requirement on the data for unmap
 - ❑ If “RZ” device, data must be all zeros
 - ❑ If not “RZ” device, can’t know the data pattern
 - ❑ Conclusion: WRITE SAME with UNMAP bit is usable only if LBPRZ = 1

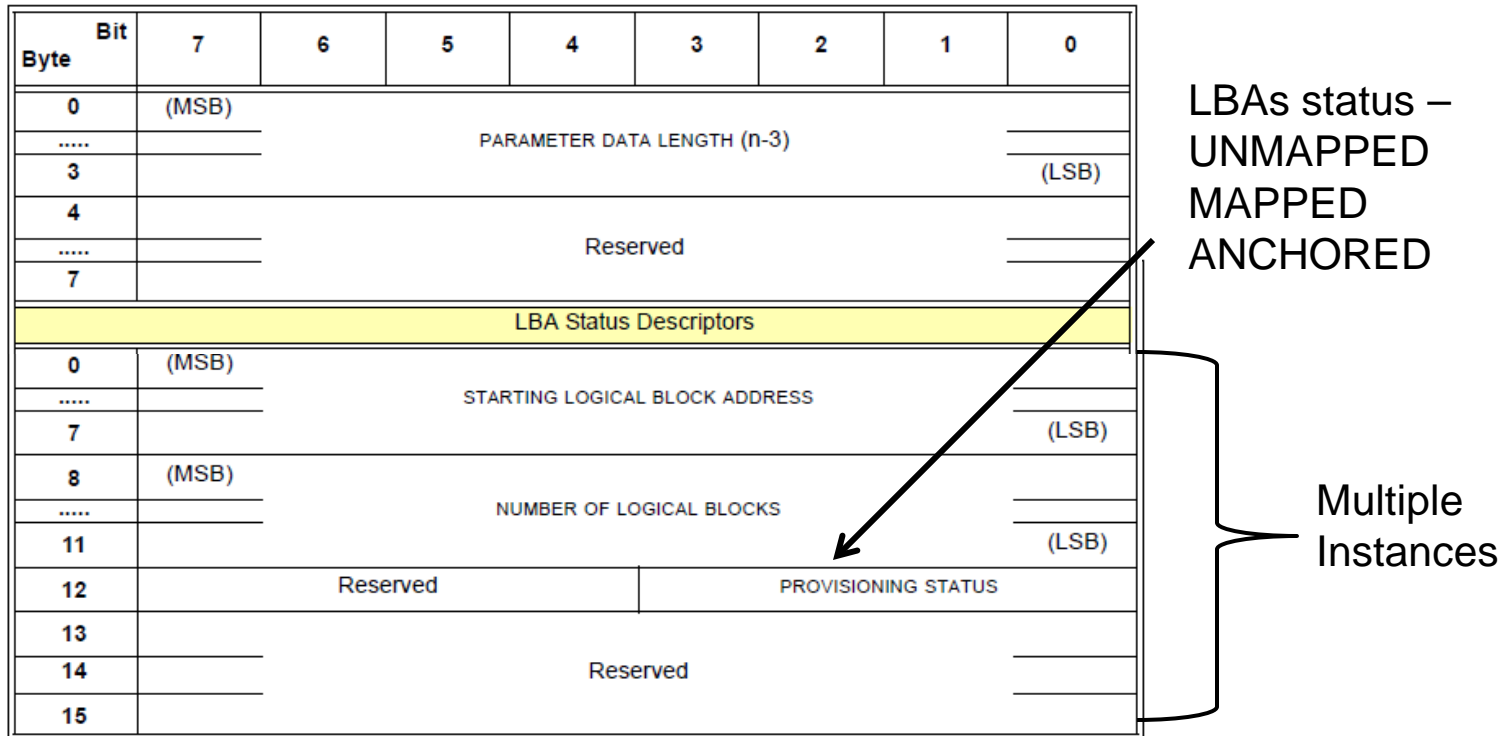
SCSI Cmds – GET LBA STATUS

- ❑ Query status of LBA (Does LBA contain data?)
 - ❑ Volume backup of valid data
 - ❑ Create mirror from valid data
 - ❑ Background unmapped reclaimer

Byte	Bit	7	6	5	4	3	2	1	0
0		OPERATION CODE (9Eh)							
1		Reserved			SERVICE ACTION (12h)				
2	(MSB)								
.....		STARTING LOGICAL BLOCK ADDRESS							
9		(LSB)							
10	(MSB)								
.....		ALLOCATION LENGTH							
13		(LSB)							
14		Reserved							
15		CONTROL							

SCSI Cmds – GET LBA STATUS

□ Extent list with status

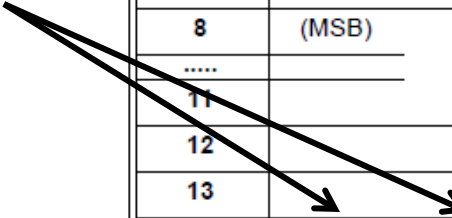


- ❑ Discovery (is my LUN a Thin provisioned)?
 - ❑ REPORT SUPPORTED OPERATION CODES
 - ❑ UNMAP command with zero xfer length
 - ❑ READ CAPACITY (16) returned data
 - ❑ VPD Page B2h returned data
 - ❑ LBPU bit and LBPWS bits
 - ❑ Limits and granularity

SCSI Commands – Parameters

- ❑ Discovery (is my LUN a Thin provisioned)?
 - ❑ READ CAPACITY (16) returned data

Byte	Bit	7	6	5	4	3	2	1	0	
0	(MSB)	RETURNED LOGICAL BLOCK ADDRESS								
.....										
7										(LSB)
8	(MSB)	LOGICAL BLOCK LENGTH IN BYTES								
.....										
11										(LSB)
12		Reserved				P_TYPE		PROT_EN		
13		P_I_EXPONENT				LOGICAL BLOCKS PER PHYSICAL BLOCK EXPONENT				
14		LBPME	LBPRZ	(MSB)	LOWEST ALIGNED LOGICAL BLOCK ADDRESS					
15		(LSB)								
16		Reserved								
.....										
31										



SCSI Commands – Parameters

- Discovery (is my LUN Thin provisioned)?
 - VPD page B2h returned data

Byte	Bit	7	6	5	4	3	2	1	0
0		PERIPHERAL QUALIFIER			PERIPHERAL DEVICE TYPE				
1		PAGE CODE (B2h)							
2	(MSB)	PAGE LENGTH (n-3)							
3									
4		THRESHOLD EXPONENT							
5		LBPW	LBPWS	LBPWS10	Reserved		<u>LBPRZ</u>	ANC_SUP	DP
6		Reserved					<u>PROVISIONING TYPE</u>		
7		Reserved							
8		PROVISIONING GROUP DESCRIPTOR							
.....									
n									

Unreported
Resource
Thin

SCSI Commands – Parameters

- ❑ Discovery (Limits and Granularity)?
 - ❑ VPD page B0h returned data

19			(LSB)
20	(MSB)		
.....		MAXIMUM UNMAP LBA COUNT	
23			(LSB)
24	(MSB)		
.....		MAXIMUM UNMAP BLOCK DESCRIPTOR COUNT	
27			(LSB)
28	(MSB)		
.....		OPTIMAL UNMAP GRANULARITY	
31			(LSB)
32	UGAVALID	(MSB)	
.....		UNMAP GRANULARITY ALIGNMENT	
35			(LSB)
36	(MSB)		
.....		MAXIMUM WRITE SAME LENGTH	
43			(LSB)
44			

SCSI Commands – Parameters

Resource count reporting

Log page 0Ch

Byte	Bit	7	6	5	4	3	2	1	0
0		DS (1)	SPF (0)	PAGE CODE (0Ch)					
1		SUBPAGE CODE (00h)							
2	(MSB)	PAGE LENGTH (n - 3)							
3		(LSB)							
Logical Block Provisioning parameter list									
0	(MSB)	PARAMETER CODE (0001h)							
1		(LSB)							
2		Parameter control byte – binary format list log parameter (see SPC-4)							
		DU	Obsolete	TSD	ETC	TMC	FORMAT AND LINKING		
3		PARAMETER LENGTH (0408h)							
4	(MSB)	RESOURCE COUNT							
.....		(LSB)							
7									
8		Reserved						SCOPE	
9									

Available
Used
Dedup
Compressed

Multiple
Instances

1 LUN only
> 1 LUN

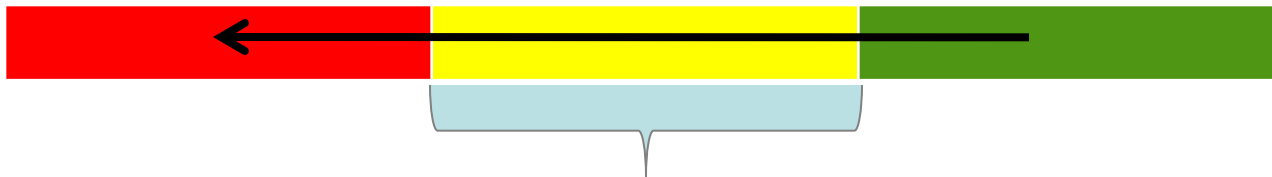
SCSI Resource Thresholding

- Threshold zones (yellow zone)

- Armed Increasing Threshold - used resources



- Armed Decreasing Threshold - available resources

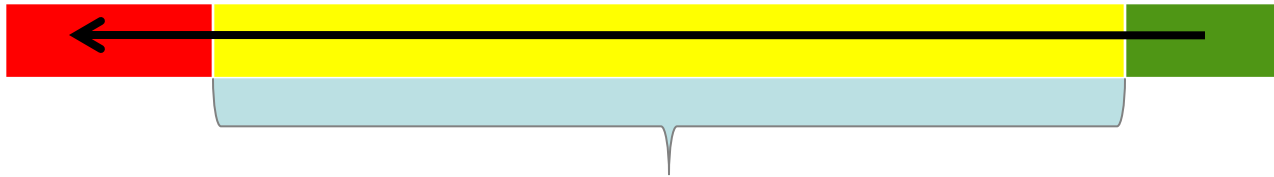


Unit Attention

- UA – THIN PROVISIONING SOFT THRESHOLD REACHED

SCSI Resource Thresholding

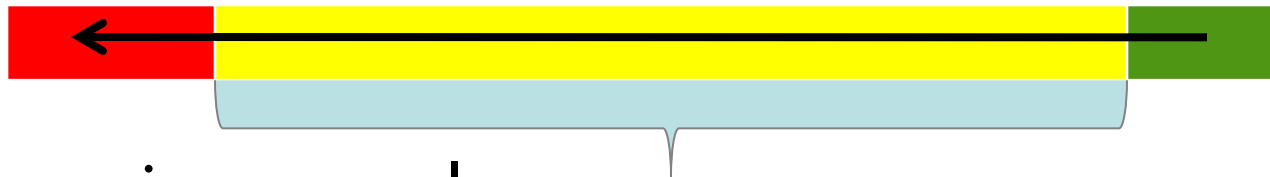
- Threshold zones (yellow zone)



- Set at the center point of each yellow zone
 - Blocks based on a power of 2 (threshold set size)
 - 10 = 1024 blocks, 12 = 4096 blocks, 16 = 64k blocks
- UA arms on entry
- UA triggers at or before exiting

SCSI Resource Thresholding

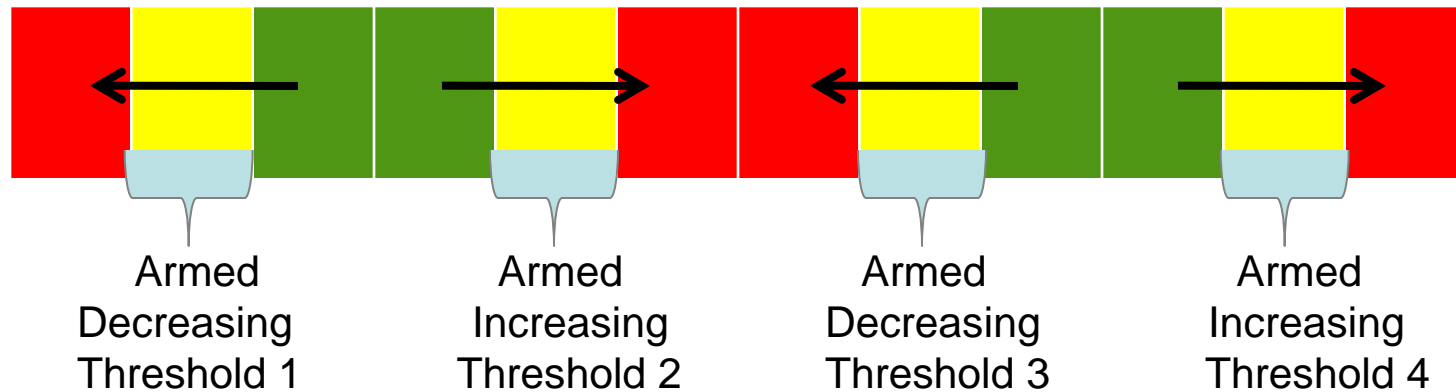
- Threshold zones (yellow zone)



- Decreasing example:
 - Threshold set size = 10 (1024 blocks)
 - Threshold count = 16 (16384 blocks)
 - As the available space decreases, we arm the UA at 16896 (16384 + 512) blocks and deliver the UA at or before the available space decreases to 15872 (16384 – 512) blocks

SCSI Resource Thresholding

- Multiple thresholds may exist



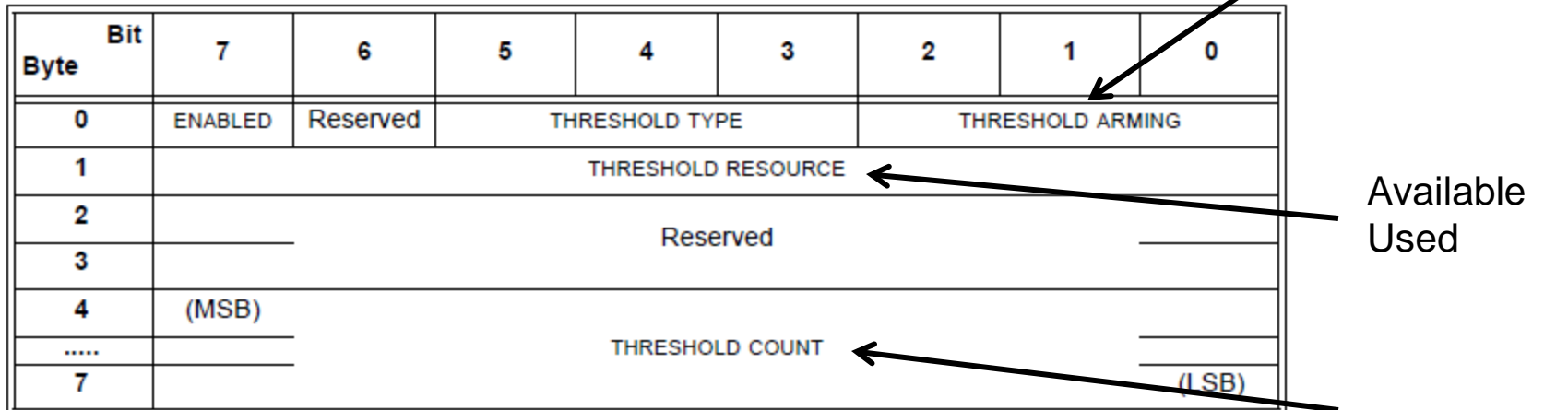
- Each threshold is set at the center of it's yellow zone
- When encountered, the UA sense data will indicate which threshold was triggered

SCSI Resource Thresholding

- ❑ Uses standard MODE SENSE/SELECT model
- ❑ MODE SENSE (get current page)
- ❑ MODE SENSE (get changeable page)
- ❑ If none are changeable, then threshold are hard set by the storage vendor
- ❑ The changeable fields within a threshold can be set by host (maybe the set point, maybe the arming direction)

SCSI Resource Thresholding

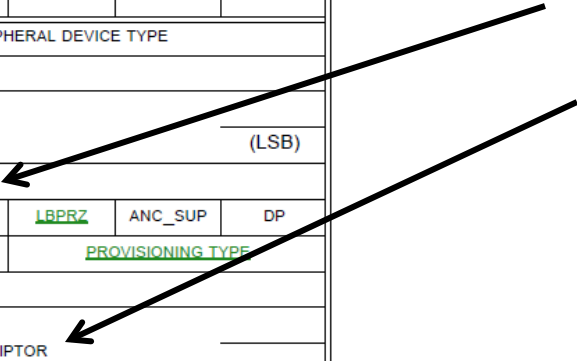
- Resource threshold reporting
 - Mode page I Ch subpage 02h
 - Multiple threshold descriptors



SCSI Resource Thresholding

- ❑ VPD page defines the width of threshold (the yellow zone) defined by device vendor (not changeable)
- ❑ VPD page contains a Provisioning Group Descriptor
 - ❑ LUNs with the same descriptor share the same resource pool

Byte	Bit	7	6	5	4	3	2	1	0
0	PERIPHERAL QUALIFIER			PERIPHERAL DEVICE TYPE					
1	PAGE CODE (B2h)								
2	(MSB)	PAGE LENGTH (n-3)							(LSB)
3	PAGE LENGTH (n-3)								
4	THRESHOLD EXPONENT								
5	LBPU	LBPWS	LBPWS10	Reserved		LBPRZ	ANC_SUP	DP	
6	Reserved					PROVISIONING TYPE			
7	Reserved								
8	PROVISIONING GROUP DESCRIPTOR								
.....	PROVISIONING GROUP DESCRIPTOR								
n	PROVISIONING GROUP DESCRIPTOR								



- ❑ DATA SET MANAGEMENT command
 - ❑ TRIM function (word 69 bit 0)
 - ❑ Supplies a list of LBA Range Entries
 - ❑ 48 Bit LBA number and 16 bit length
 - ❑ DRAT vs. non-DRAT (word 69 bit 14)
 - ❑ RZAT vs. non-RZAT (word 69 bit 5)

ATA & SATA Commands

- ❑ Issue DSM command with Trim = 1
 - ❑ No Write data is transferred
 - ❑ Supply list of Starting LBA + Length in buffer
 - ❑ Therefore SCSI UNMAP is an easy translation

- ❑ Deleting a file
 - ❑ UNMAP the files LBAs
- ❑ Background scanner
 - ❑ GET LBA STATUS
 - ❑ READ the file system allocation bit map
 - ❑ UNMAP the LBAs that are not allocated
- ❑ Mirror Creation
 - ❑ GET LBA STATUS
 - ❑ Copy only real data

- ❑ Learn what File System APIs are available
- ❑ Ask Vendors to provide APIs