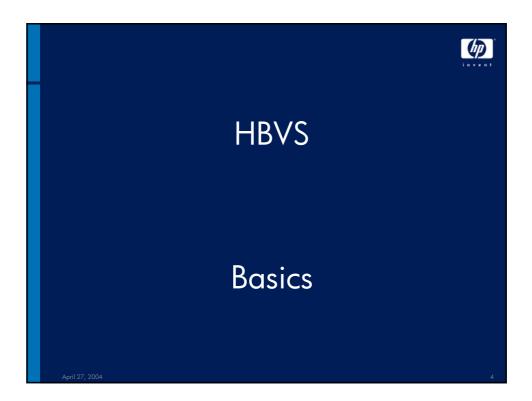


Topics • Features post V7.3-2 -Merge and Copy Prioritization -Copy or Merge Host System Selection -Host Based Mini Merge (HBMM) • Long Distance RAID1





HBVS Basics

- Shadow set (a.k.a. virtual unit or VU) normally consists of multiple shadow set member (SSM) units
- Application Write I/O is sent to all SSMs
 - In parallel to all full members
 - Then to all copy members
- Application Read I/O is done from a "source" (a.k.a. full) member
 - Uses the individual SSM "read cost" and queue depth

April 27, 2004

5

Why is a Merge Needed



- A system has a VU mounted write enabled
 - If that system crashes
 - 01
 - If that system aborts Mount Verification on that VU, with write I/O in an internal restart queue
- Then
 - Write I/O "in-flight" state is indeterminate
 - All, some, or none of the SSMs may have been written
 - Application read I/O could have the potential to read different data for the same block on different SSMs
- Remaining systems have no inherent knowledge about application write I/O state at that point

April 27, 2004

6



Why is a Merge Needed

Every application read I/O must be merged.

- Merge operation will
 - Read and compare extent on all members
 - Fix differences found
 - Return the read to application
- Recall that application write I/O was never acknowledged as having completed
 - Therefore there is no "correct" data ... only consistent data

April 27, 2004

7

Why is a Merge Needed



- Shadow Server process is used to insure that all the blocks of a volume get merged
- SHADOW_MAX_COPY determines number of concurrent threads any Shadow Server can run
- This merge operation maintains a "merged fence"
 - Fence starts at LBN 0 of the volume
 - Blocks below the fence are considered merged
 - Blocks above the fence are considered unmerged

April 27, 2004

8



DVE

Dynamic Volume Expansion

April 27, 200

(p)

Why Dynamic Volume Expansion

- Need to grow volume sizes with minimal impact on operations
- New controllers can expand the size of a devices without taking it off-line
- Host Based Volume Shadowing (HBVS) enables volume growth via Dissimilar Device Support

April 27, 2004

10



DVE - Preparation

- Determine / decide how big this volume may ever get to make maximum use of the dynamic expansion capability
- Consider using the maximum ... 1 TB
 - Modest file system overhead cost (32MB) in disk space
 - Will allow dynamic online growth in the future

April 27, 200

11

How DVE works



- New term: logical volume size
- New DCL commands to create a storage bitmap that is large enough for future growth
- Two options
 - For new volumes use
 - \$INIT with new command qualifiers
 - For existing volumes use
 - SET VOLUME /LIMIT to expand its potential size
 - -This **requires** volume to be mounted privately

April 27, 2004

12



INITIALIZE Qualifiers for DVE **S INITIALIZE /SIZE**

- \$ INITIALIZE / SIZE
- Sets the current Logical Volume Size (i.e. SCB\$L_VOLSIZE)
 of the volume
- Defaults to UCB\$L_MAXBLOCK of the device
- Can be made less than UCB\$L_MAXBLOCK

\$ INITIALIZE /LIMIT

- Sets the maximum growth size, i.e. generates a bitmap that will support this limit
 - Default is 1 Terabyte (suggested)
- Rounds off the expansion size to the use the full bitmap block

NOTE: If /LIMIT is used, the default /CLUSTER will be 8

April 27, 2004

12

SET VOLUME Qualifiers for DVE



- SET VOLUME/LIMIT
 - Must be done while MOUNTed privately
 - Not /SYSTEM or /CLUSTER
 - Prepares a volume for future expansion by extending or moving the bitmap as needed
 - Does not change clustersize
 - Does not change the logical volume size
 - Rounds off the expansion size to use the full bitmap block

April 27, 2004

14

SET VOLUME Qualifiers for DVE (continued)



· SET VOLUME/SIZE

- Extends the current logical volume to the size specified
- May be done online with applications active
- Increases SCB\$L VOLSIZE
- Will not reduce the current size of the mounted volume
- Will not extend beyond UCB\$L_MAXBLOCK or the capacity of the storage bitmap

April 27, 2004

15

Additional Information on Volume Characteristic



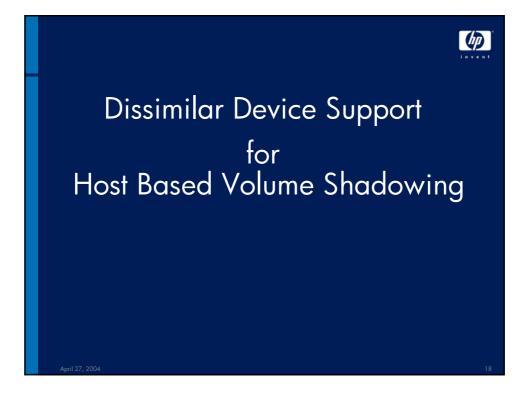
Show Device/Full

- "Total Blocks" reports actual size of device or storage container (derived from UCB\$L_MAXBLOCK)
- Additional display information reports current
 - logical volume size
 - expansion size limit
- SGETDVI programming and F\$GETDVI lexical support for new fields

April 27, 2004

16







Why Dissimilar Device Support?

Flexibility and lower cost

- Drives can have variations in total blocks
 - Controller variations can cause total blocks to be different, using the same physical device
- Granularity in control of size in "virtualizing" controllers does not exist
- Consolidation of existing devices
 - Enables CI or local SCSI devices to be shadowed with FC devices

April 27, 2004

10

How it works



- Selection of Founding Member remains the same
 - When shadow set is initially created -
 - the "founding device" is the de facto shadow set master member
- Founding Member SCB\$L_VOLSIZE is the minimum size of incoming shadow set members
- No change in MOUNT interface
- New shadow set members (copy targets) must have at least SCB\$L_VOLSIZE blocks to be added to the virtual unit

April 27, 2004

20

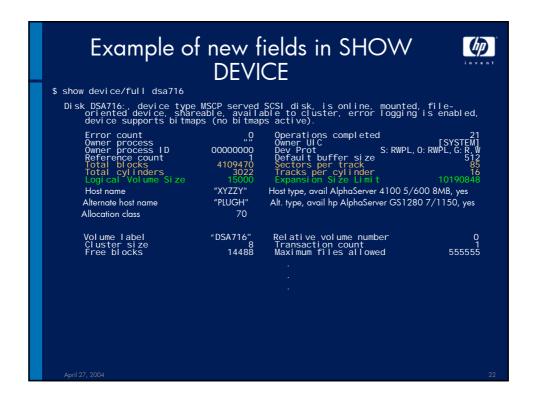
How it works



- Virtual Unit UCB\$L_MAXBLOCK maintained as that of the smallest shadow set member
 - The geometry (sectors / tracks / cylinders) of the virtual unit will be maintained to the smallest shadow set member
 - This geometry information not used by HBVS

April 27, 200

21



DDS and DVE



Putting these two features together means that taking a volume offline to increase its capacity or size is no longer necessary, once the limit has been set

- Use the new command qualifiers for
 - \$INITIALIZE
 - SET VOLUME device:
 - With a very large bitmap (/LIMIT) expansion

April 27, 200

23

DDS and DVE



- If volume is mounted as a single shadow set member
 - When more space is needed, add a larger physical device and wait for the copy operation to complete
 - Remove the smaller member
 - Now there is room to expand the volume
 - Expand the volume (SET VOLUME/SIZE)

Repeat as needed

April 27, 2004

24

DVE / DDS Availability



• DDS and DVE is available in V7.3-2

• DDS will also be available in the HBMM kit

April 27, 200

25

SET / SHOW SHADOW

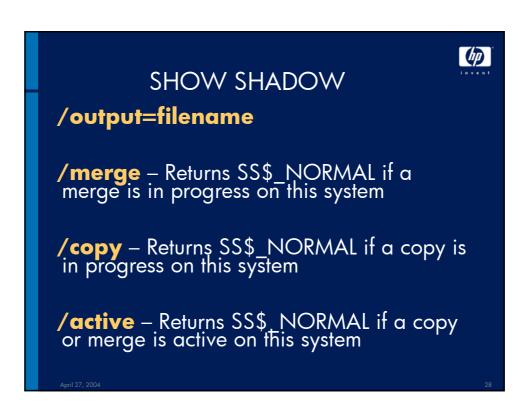


- New utility in V7.3-2
 Will be available on V7.3-1 with HBMM kit
- SET SHADOW
- SHOW SHADOW
- ANALYZE/DISK/SHADOW

April 27, 2004

26





SET SHADOW



/output=filename – outputs any messages to the specified file

/log – display a brief message that confirms that the command completed

April 27, 200

20

SET SHADOW (cont.)



/site – sets the site value for the VU only, use SET DEVICE/SITE for members

/mvtimeout – sets mvtimeout for VU

/abort_virtual_unit - Causes an immediate abort of MountVerification on the virtual unit

April 27, 2004

30



ANALYZE / DISK / SHADOW

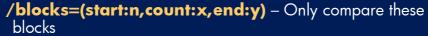
- Used to verify that all full members, not copy members, have the same information on all blocks
- Earlier compare utilities could get "transient" miss compares, if application "hot blocks" were encountered
 - This utility eliminates transient miss compares because the VU is write locked, the blocks are re-compared, and only then is a problem reported
- File name is displayed and the actual data block is dumped

April 27, 2004

31

伽

ANALYZE /DISK /SHADOW DSAn:



/brief – Displays only the LBN if a difference is found. Without this qualifier, if the LBN has differences, the LBN on all members is dumped to the screen

/file_system – Only report errors if the LBN is within the file system

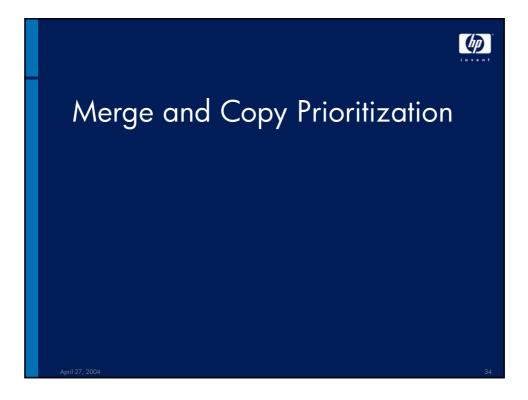
/ignore – Ignore 'special' files – i.e. SYSDUMP

/output=filename – output the information to the specified file

/statistics – only display the header and summary statistics

April 27, 2004

32



Shadow Set State Hierarchy • Mini Merge state • Copy state • Mini Copy state • Full Copy state • Full Merge state • Steady state

Current Merge and Copy Controls



- Management controls to determine order and choice of system for copy and merge operations are incomplete.
- Important volumes may be merged after less important volumes
- Systems better suited to perform merge or copy operations on some volumes are not always selected

April 27, 2004

36



Merge and Copy Control

- Allow user to assign a priority to every VU
- Better predict which system will perform any transient state operations (merge or copy operations) – requires SYSGEN settings
- Utilize SHADOW_MAX_COPY dynamic characteristic

April 27, 200

37

Shadow Priority



- New command qualifier
 - \$ SET SHADOW PRIORITY = n DSAnnnn:
 - A range of 0 through 10,000
 - Default is 5000
 - 1 is the lowest priority
 - Zero has special meaning
- At MOUNT time each VU will be placed in system wide priority linked list by this value
 - VUs at the same priority have an undefined ordering
- Governs merge and copy priority for VUs on this system

April 27, 2004

38



Show Shadow Priority

New command

\$ SHOW SHADOW /BY_PRIORITY

- Lists the DSA devices on this system using the priority assigned to each, highest to lowest
- Shows transient state % and system performing operation

Device Priority Virtual Unit State % Completed on

Node

DSA3233: 3233 Steady State
DSA2325: 2325 Not Mounted

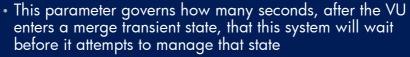
DSA42: 42 Full Merge Active 14% on ATHRUZ

April 27, 2004

39

New SYSGEN parameter

SHADOW_REC_DLY (Shadow Recovery Delay)

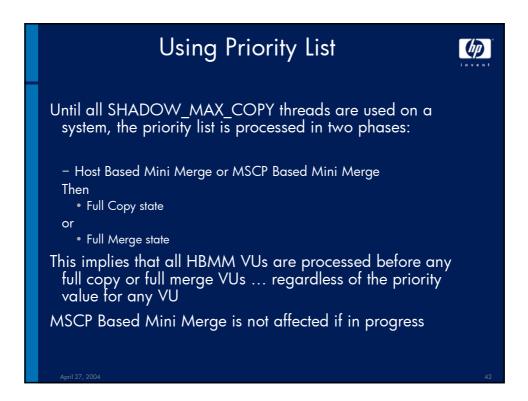


- RECNXINTERVAL is added to the total wait time
- Default of 20 seconds
- Making this value different across the cluster will guide which system will manage transient state operations on which VU

April 27, 2004

40

In Hierarchal order: - Mini Merge state • Host Based Mini Merge • MSCP Based Mini Merge - Also known as write logging - Copy state • is a Shadow Set Member (SSM) Specific - Mini Copy state - Full Copy state - Full Merge state



Managing Transient State Operations



New SET SHADOW command qualifiers:

/EVALUATE=RESOURCES

- Is a system specific command ... i.e. not cluster specific
 - Thus it only affects the VUs that are being merged or copied by this system
- Useful if the command is issued before
 - SHADOW_MAX_COPY is increased or is decreased
 - -To start or stop a merge or a copy operation
 - The priority of a VU is changed
 - Priority of 0 now non-zero or the inverse
 - Priority value has been raised or lowered

April 27, 2004

43

Managing Transient State Operations



New SET SHADOW command qualifier:

/DEMAND_MERGE DSAn:

Changes the state of the VU to a merge required state

- The type of merge initiated will depend on merge recovery characteristics that are enabled currently on the VU
- To insure that a full merge is initiated on a VU
 - Disable any mini merge that is currently enabled
- This is especially useful if the shadow set had been created with INIT/ SHADOW without / ERASE
- Can be used if differences are found on the members with ANALYZE /DISK /SHADOW

April 27, 2004

44



What is Mini Merge?



- Full merge requires comparing entire shadow set
- But only blocks with I/O in progress need to be merged
- MSCP Based mini merge
 - Supported on HSCxx / HSDxx / HSJxx controllers
 - Controller tracks in-progress writes
 - Host can get list of writes from the controller
- Host Based Mini Merge (HBMM)
 - Selected cluster hosts track recent writes using write bitmap
 - Bitmap is periodically reset to flush out old writes
 - Contents of bitmap drive mini merge operation

April 27, 2004

46



Write Bitmap for Mini Merge

- Write bitmap originally released on V7.2-2 for mini copy
 - Each system has an in-memory bitmap to track write I/O
 Each bitmap has one system that is the master
 2KB memory per Gbyte of storage per bitmap per system
- There are 6 bitmaps are available (per VU) for mini merge use
 - There are also 6 bitmap slots reserved for mini copy use
- HBMM recovery must be by a bitmap master system
- Use multiple bitmap masters for availability of the bitmap after a system

Bitmap Master Policy



- The policy defines
 - number of bitmap masters for a VU in the cluster
 - the *location* of masters (in counted groups)
 - the bitmap reset threshold
- Named policies
 - Are known cluster wide
 - Any named policy can be assigned to an individual VU or to multiple VUs
 - A named policy can be deleted at any time
 - That does not affect VUs that had it applied



Bitmap Master Policy

- A policy must be directly assigned to VU for HBMM to be enabled on that VU
- With the following exception
 - If a DEFAULT policy has been defined, then every VU in the cluster, that does not have a named policy, will automatically "pick up" the DEFAULT policy
 - To prevent a VU from "picking up" the DEFAULT policy use
 - SET SHADOW DSA1: / POLICY = HBMM = NONE
- In summary
 - the DEFAULT policy and a mix of VU specific policies can be used in the cluster at the same time

April 27, 2004

49

Policy Definition





Other HBMM Controls

SET SHADOW DSA1: / DISABLE = HBMM

- Disables HBMM on DSA1

SET SHADOW DSA1: / ENABLE = HBMM

- Enables HBMM on DSA1
- If there is a policy for the VU to enable

SET SHADOW DSA1: / POLICY=HBMM=(MASTER=*)

- Allows all systems in the cluster to be bitmap masters
- First six to mount DSA1: will become masters
 - Others eligible when a master dismounts or crashes

Other HBMM Controls



SET SHADOW / POLICY = HBMM = (MASTER_LIST =)

/ NAME = DEFAULT

Creates default policy for all shadow sets that are mounted in the cluster that do not have a VU specific policy in place

SET SHADOW DSA1: / POLICY = HBMM = NONE - Disables HBMM for DSA1

- Can be used to override a DEFAULT policy on a specific VU
- Can be viewed as a policy governor

SET SHADOW

- / POLICY = HBMM / NAME=POLICY_1 / DELETE Will delete the policy named "POLICY_1"



HBMM rules

- If a policy is associated with a VU, HBMM is automatically enabled upon first mount on a system that has been named as bitmap master in that policy
- If a master system ceases to be a master (due to dismount or crash), a new master bitmap will be **automatically** started on another master system, subject to the policy in force on the VU
- Devices capable of MSCP Based (HSC/HSJ/HSD) mini merge are not eligible for HBMM
- To enable host based mini merge on a VU, all systems that mount the VU must be HBMM capable

April 27, 2004

55

Host Based Mini Merge Distribution



HBMM is planned to be released on:

- V7.3-1
 - On request!
- V7.3-2
 - With 8.2 Release
- No plans for VAX support

April 27, 2004

56

