


1G04


OpenVMS Shadowing
Update

Manfred Kaser
HP Services
Manfred.Kaser@hp.com



© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

Topics



- HBVS Basics
- New Features in V7.3-2/V8.2
 - Dynamic Volume Expansion
 - Dissimilar Device Support in HBVS
 - SET / SHOW SHADOW
 - ANALYZE/DISK/SHADOW

April 15, 2005

2

Topics



- Features missing post V7.3-2
 - Merge and Copy Prioritization
 - Copy or Merge Host System Selection
 - Host Based Mini Merge (HBMM)

April 15, 2005

3

HBVS

Basics



April 15, 2005

4



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 15, 2005

5



Why is a Merge Needed

- A system has a VU mounted write enabled
 - If that system crashes
 - or
 - 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 15, 2005

6



Why is a Merge Needed cont.

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 15, 2005

7




Why is a Merge Needed cont.

- 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 15, 2005


8



DVE

Dynamic Volume Expansion

April 15, 2005 9



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 15, 2005 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 15, 2005

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 15, 2005

12



INITIALIZE Qualifiers for DVE

- **\$ 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 use the full bitmap block

Q R WH #L#2CP I#V#xhg/#h#ghid#2F OX VWHU#z lo#h#; #

April 15, 2005

13



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 15, 2005

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 15, 2005

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**
 - \$GETDVI programming and F\$GETDVI lexical support for new fields

April 15, 2005


16



\$ SHOW DEVICE/FULL Extract

Total blocks	4109470	Sectors per track	85
Total cylinders	3022	Tracks per cylinder	16
Logical Volume Size	15000	Expansion Size Limit	2147483647
Host name	"XYZZY"	Host type, avail	Alpha ...

April 15, 2005 17



Dissimilar Device Support for Host Based Volume Shadowing

April 15, 2005 18



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 15, 2005

19



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 15, 2005

20



How it works

- Virtual Unit UCB\$_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 15, 2005

21

Example of new fields in SHOW DEVICE



```
$ show device/full dsa716
Disk DSA716:, device type MSCSP served SCSI disk, is online, mounted, file-
oriented device, shareable, available to cluster, error logging is enabled,
device supports bmaps (no bmaps active).

Error count                "0"          Operations completed          21
Owner process              "0"          Owner UIC                    [SYSTEM]
Owner process ID          00000000   Dev Prot                     S: RWPL, O: RWPL, G: R, W
Reference count            1          Default buffer size          512
Total blocks               4109470   Sectors per track            85
Total cylinders            3022     Tracks per cylinder           16
Logical Volume Size       15000     Expansion Size Limit         10190848
Host name                  "XYZZY"    Host type, avail AlphaServer 4100 5/600 8MB, yes
Alternate host name        "PLUGH"    Alt. type, avail hp AlphaServer GS1280 7/1150, yes
Allocation class           70

Volume label               "DSA716"   Relative volume number       0
Cluster size               8          Transaction count            1
Free blocks                 14488     Maximum files allowed        555555
.
.
.
```

April 15, 2005

22

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 15, 2005

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 15, 2005

24

DVE / DDS Availability



- DDS and DVE is available in $\geq V7.3-2$
- DDS will also be available in the HBMM kit

April 15, 2005

25

SET / SHOW SHADOW




- New utility in V7.3-2
- SET SHADOW
- SHOW SHADOW
- ANALYZE/DISK/SHADOW

April 15, 2005

26

SHOW SHADOW



_DSA3233: DSA3233 - Shadowing Level 2 In Use
Virtual Unit State: Steady State


VU Timeout Value	3600	VU Site Value	2
Copy/Merge Priority	3233	Mini Merge	Enabled

HBMM Policy
HBMM Reset Threshold: 17777865 blocks
HBMM Master lists:
Up to any 2 of the nodes: ATHRUZ,ATWOZ
Modified blocks since bitmap creation: 0

Device \$1\$DGA32			
Read Cost	2	Site 1	
Member Timeout	120		
Device \$1\$DGA33		Master Member	
Read Cost	2	Site 2	
Member Timeout	120		

April 15, 2005 27

SHOW SHADOW



/output=filename

/merge – Returns SS\$_NORMAL if a merge is in progress on this system

/copy – Returns SS\$_NORMAL if a copy is in progress on this system

/active – Returns SS\$_NORMAL if a copy or merge is active on this system

April 15, 2005 28



SET SHADOW

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

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

April 15, 2005

29



SET SHADOW (cont.)

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

/mvertimeout – sets mvertimeout for VU

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

April 15, 2005

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 15, 2005

31




ANALYZE /DISK /SHADOW DSA_q:

- /blocks=(start:n,count:x,end:y)** – Only compare these blocks
- /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 15, 2005

32



ANALYZE /DISK /SHADOW output

```
$ anal /di sk/shadow/bri ef/bl ock=count=1000 dsa716:

Starting to check _DSA716: at 14-MAY-2003 13:07:47.01
Members of shadow set _DSA716: are _$252$MDAO: _$252$DUA716:
and the number of blocks to be compared is 1000.
Checking LBN #0 (approx 0%)
Checking LBN #127 (approx 12%)
Checking LBN #254 (approx 25 %)
Checking LBN #381 (approx 38%)
Checking LBN #508 (approx 50%)
Checking LBN #635 (approx 63%)
Checking LBN #762 (approx 76%)
Checking LBN #889 (approx 88%)


Run statistics for _DSA716: are as follows:
Finish Time = 14-MAY-2003 13:07:47.30
ELAPSED TIME = 0 00:00:00.29
CPU TIME = 0:00:00.02
BUFFERED I/O COUNT = 10
DIRECT I/O COUNT = 16
Failed LBNs = 0
Transient LBN compare errors = 0
```

April 15, 2005 33



Merge and Copy Prioritization

April 15, 2005 34




Shadow Set State Hierarchy

- Mini Merge state
- Copy state
 - Mini Copy state
 - Full Copy state
- Full Merge state
- Steady state

Transient states

April 15, 2005 35



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 15, 2005 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 15, 2005

37



Shadow Priority

- New command qualifier
 - \$ SET SHADOW /PRIORITY = n DSAqqrqr:
 - 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 15, 2005

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 Node	Priority	Virtual Unit State	% Completed on
DSA3233:	3233	Steady State	
DSA2325:	2325	Not Mounted	
DSA42:	42	Full Merge Active	14% on ATHRUZ

April 15, 2005

39



New SYSGEN parameter

SHADOW_REC_DLY (Shadow Recovery Delay)

- This parameter governs how many seconds, after the VU enters a merge transient state, that this system will wait before it attempts to manage that state
 - REC_NXINTERVAL 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 15, 2005

40

Possible Transient States



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

April 15, 2005

41

Using Priority List



Until all SHADOW_MAX_COPY threads are used on a system, the priority list is processed in two phases:

- Host Based Mini Merge or MSCP Based Mini Merge
- Then
- Full Copy state
- or
- Full Merge state

This implies that all HBMM VUs are processed before any full copy or full merge VUs ... regardless of the priority value for any VU

MSCP Based Mini Merge is not affected if in progress

April 15, 2005

42

Managing Transient State Operations

New SET SHADOW command qualifiers:
/EVALUATE=RESOURCES

- Is a v|w|hp 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 15, 2005 43

Managing Transient State Operations

New SET SHADOW command qualifier:
/DEMAND_MERGE DSA:

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 15, 2005 44



Host Based Mini Merge

April 15, 2005 45



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 15, 2005 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 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 crash

April 15, 2005

47



Bitmap Master Policy

- The policy defines
 - the number of bitmap masters for a VU in the cluster
 - the ordering 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

April 15, 2005

48



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 15, 2005

49



Policy Definition

```
SET SHADOW/POLICY = HBMM =
  ( (MASTER_LIST = (NODE1,NODE2,NODE3),
    COUNT=2),
    (MASTER_LIST = (NODE4,NODE5,NODE6),
    COUNT=2),
    RESET_THRESHOLD=50000) / NAME =POLICY_1
```

```
SET SHADOW DSA1: / POLICY = HBMM = POLICY_1
```

April 15, 2005

50



Policy Definition continued

```
SET SHADOW DSA1: /POLICY = HBMM =  
  ( (MASTER_LIST = (NODE1,NODE2,NODE3),  
    COUNT=2),  
    (MASTER_LIST = (NODE4,NODE5,NODE6),  
    COUNT=2),  
    RESET_THRESHOLD=50000)
```

April 15, 2005

51



Policy Usage

```
MOUNT/CLUSTER  
  DSA1: / SHADOW=($1$DGA22:,$1$DGA23:)  
  LABEL
```

April 15, 2005

52



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

April 15, 2005

53



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"

April 15, 2005


54



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 p r x q w the VU must be HBMM capable

April 15, 2005 55



Host Based Mini Merge Distribution

HBMM is available

- V7.3-2
plus VMS732_UPDATE-V0300
- V8.2
- No VAX support

April 15, 2005 56

Fragen ?

