



GET CONNECTED

PEOPLE. TECHNOLOGY. RESULTS.

Using Shadowsets With More Than 3 Members

Keith Parris

Systems/Software Engineer, HP

June 16, 2009

Session 3033



Background

- Shadowsets have been limited to a maximum of 3 shadowset members
- OpenVMS Version 8.4 is slated to support up to 6 shadowset members:

```
DS10 $ show dev dsa5678:
```

Device Name	Device Status	Error Count	Volume Label	Free Blocks	Trans Count	Mnt Cnt
DSA5678:	Mounted	0	SIXMEMBER	682944	1	1
\$6\$DKB0:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			
\$6\$DKB100:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			
\$6\$DKB200:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			
\$6\$DKB300:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			
\$6\$DKB400:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			
\$6\$DKB500:	(WSC236) ShadowSetMember	0	(member of DSA5678:)			

```
DS10 $
```



Why Have Customers Asked HP to Raise the Limit?



Why Have Customers Asked HP to Raise the Limit?

- Disaster-Tolerant OpenVMS Clusters:
 - Retain redundancy after loss of 1 of 2 sites
 - Some customers have 2-site DT clusters and can survive loss of 1 site and continue processing with no data loss
 - In such a 2-site cluster, one site could have 2 shadowset members, but the other site could only have 1 shadowset member. If you lost the site with 2 shadowset members, you were down to only a single copy of the data, and thus vulnerable until a shadow-copy adds a 2nd member at that site
 - 4 shadowset members allows 2 shadowset members per site

Why Have Customers Asked HP to Raise the Limit?

- Disaster-Tolerant OpenVMS Clusters:
 - Retain redundancy after loss of 2 of 3 sites
 - Some customers have 3-site DT clusters and can survive loss of any two of the 3 sites and continue processing, uninterrupted, and with no data loss.
 - In a 3-site cluster, each of the sites can only have 1 shadowset member. If you lose one site, you are still protected, but if you lost 2 sites, you were down to only a single copy of the data.
 - 6 members allows 2 shadowset members per site

Why Have Customers Asked HP to Raise the Limit?

- Disaster-Tolerant OpenVMS Clusters:
 - Historical Work-arounds:
 - Controller-based mirroring at 1-member site
 - Controller [pair] is still a single point of failure
 - Alternate the 3rd member between sites. Remove 2nd member at one site temporarily, put a 2nd member in temporarily at the other site, and then dismount it with /POLICY=MINICOPY to keep a Mini-Copy Write Bitmap around for quick addition of a 2nd member after loss of the other site

Other Advantages of More Than 3 Shadowset Members

- More flexibility:
 - Bring extra shadowset members in temporarily and pull them out with Mini-Copy Write Bitmaps
 - Point-in-Time Copies of Data:
 - Offline backups
 - Data snapshots
 - Perhaps do this simultaneously at 2 or 3 different sites
 - Migrate existing 3-member shadowsets to new storage technology, without downtime, by temporarily forming a 6-member shadowset

Other Advantages of More Than 3 Shadowset Members

- Faster Access to Read-Only Data:
 - 3-member shadowset allowed 3 times as many reads per second as a single disk LUN
 - 6-member shadowset allows 6 times as many reads per second to the same data

Why Add This Feature Now?

Enablers For This Feature:

- Host-Based Mini-Merge
 - Minimizes performance impact of reads ahead of the merge fence with larger shadowset member counts
- Larger Lock Value Blocks (64 bytes vs. 16 bytes) in Version 8.2 and above
 - Support for 64-byte Lock Value Blocks allows passing 16 longwords between nodes (e.g. copy/merge fence, one longword per shadowset member)

Changes Required to Support More Than 3 Members



Changes Required to Support More Than 3 Members

- SHDRIVER
- Shadow Server
- \$MOUNT and \$INITIALIZE
- \$SHOW and \$SET
- \$ANALYZE/DISK/SHADOW
- SDA
- \$GETDVI, F\$GETDVI
- EXCEPTION.EXE
- Other Software:
 - HP RAID Software for OpenVMS
 - Availability Manager



Changes Required to Support More Than 3 Members

- Storage Control Block layout
 - Design had only 3 slots for shadowset device status and member IDs
 - Design now has slots for up to 16 members
 - 3 in original location
 - Array of 16 more in new location (the first 3 of which are unused)

Viewing SCB Layout with DISKBLOCK

```
DISKBLOCK> select dsa5678:/override
DISKBLOCK> read/scb
%DSKB-I-BLKREAD, Block 8886768 (%X008799F0) of _DSA5678 successfully read
DISKBLOCK> dump/scb
```

Storage Control Block

```
Structure Level:                5,1
Cluster Size:                   16
Volume Size:                    17773524 (%X010F33D4)
Blocking Factor:                1
Sectors per Track:              64 (%X00000040)
Tracks per Cylinder:            65 (%X00000041)
Number of Cylinders:            4273 (%X000010B1)
Status:                          14 (%X0000000E)
    Map Pre-Allocated
    File Numbers Pre-Allocated
    Quota file is dirty
Status2:                         0 (%X00000000)
Write Count:                     1
Volume Lock Name:                SIXMEMBER
Mount Time                      22-MAY-2009 14:33:00.63
Backup Revision Number:          0 (%X0000)
Generation Number: 12-JUN-2009 15:11:29.85
```



Viewing SCB Layout with DISKBLOCK

Virtual Unit: Type a <CR> to continue:

DSA5678:

Shadow Status: 33 (%X0021)

Shadow set populated and online

Full copy in progress

Shadow Member 0 Status: 133 (%X85)

Member involved in copy

Copy (or merge) in progress

Status information is valid

Shadow Member 1 Status: 160 (%XA0)

Member can be used for source

Status information is valid

Shadow Member 2 Status: 160 (%XA0)

Member can be used for source

Status information is valid

Member IDs:

\$6\$DKB0:

\$6\$DKB100:

\$6\$DKB200:

SCB LBN: 8886768 (%X008799F0)

Number of Devices: 6

Number of Full Members: 5

Master Index: 1

Number of Merge copy targets: 0

Number of Full Copy Targets: 1

Checksum: (Valid) Type a <CR> to continue:

22956 (%X59AC)

DISKBLOCK>



New SCB Layout

```
DS10 $ dump/hex/block=count=1 dsa5678:[000000]bitmap.sys
```

```
Dump of file DSA5678:[000000]BITMAP.SYS;1 on 12-JUN-2009 16:04:40.63  
File ID (2,2,0) End of file block 273 / Allocated 1088
```

```
Virtual block number 1 (00000001), 512 (0200) bytes
```

```
00000040 00000001 010F33D4 00100501 ....Ô3.....@... 000000  
00000000 0000000E 000010B1 00000041 A...±..... 000010  
493A2020 20524542 4D454D58 49530001 ..SIXMEMBER :I 000020  
00A8CEF8 86F081CD 000000A8 BE72ABDD Ÿ«r□¤...Í.□.øÎ¤. 000030  
000000A0 A0850021 1261162E 00000000 .....a.!..... 000040  
11620064 00000006 11620000 00000006 .....b.....d.b. 000050  
00010506 008799F0 116200C8 00000006 ....È.b.□..... 000060  
00000000 00000000 00000000 00000001 ..... 000070  
00000001 00000000 00000000 00000000 ..... 000080  
00000000 00000000 00000000 00000000 ..... 000090  
00000000 00000000 0000A0A0 A0000000 ..... 0000A0  
00000000 00000000 00000000 00000000 ..... 0000B0  
1162012C 00000006 00000000 00000000 ..... ,.b. 0000C0  
116201F4 00000006 11620190 00000006 .....b.....ô.b. 0000D0  
00000000 00000000 00000000 00000000 ..... 0000E0  
00000000 00000000 00000000 00000000 ..... 0000F0  
00000000 00000000 00000000 00000000 ..... 000100  
00000000 00000000 00000000 00000000 ..... 000110  
00000000 00000000 00000000 00000000 ..... 000120  
...  
59AC0003 00000000 00000000 00000000 .....□Y 0001F0
```



Mixed-Version Considerations

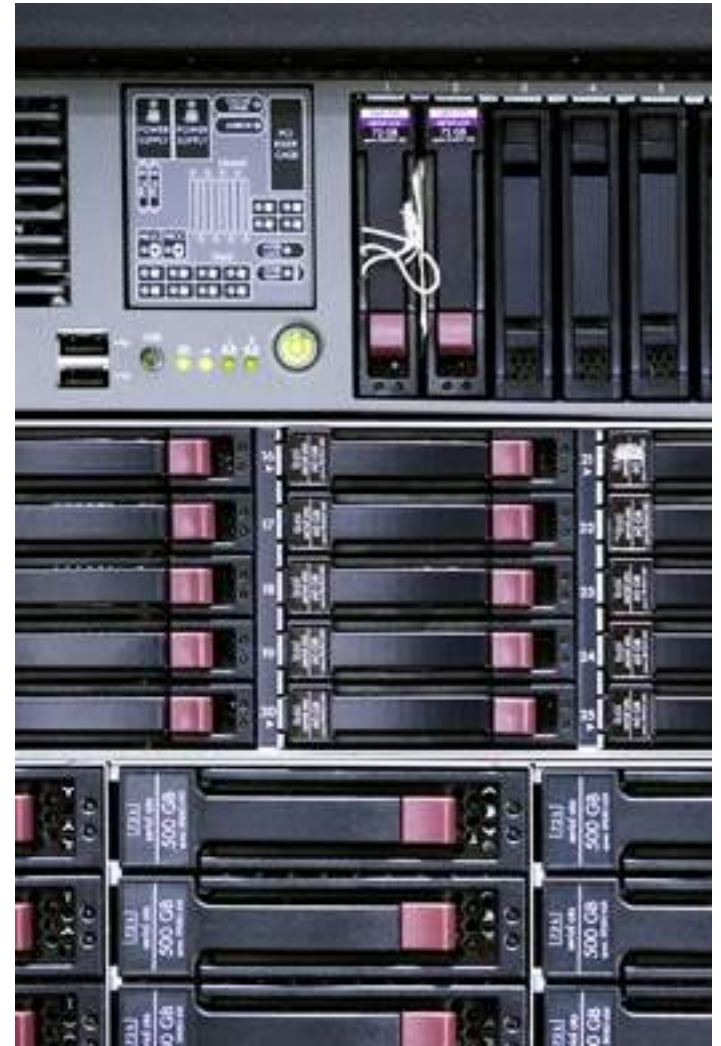


Mixed-Version Considerations

- Can co-exist in the same cluster with nodes running earlier versions than 8.4
- Support for >3 members is tracked as a per-Virtual Unit (DSA device) characteristic
- You can only mount >3-member shadowsets on nodes running Version 8.4 or above. The MOUNT command will return an error message on earlier versions.
- No plans for VAX support of >3 members



Performance Considerations



Performance Considerations

- Reads

- Potential increase in performance for reads: 6X max. vs. 3X before

Read Performance with 1-6 Members



Random 1-
block I/Os,
older local
SCSI disks,
Alphaserver
DS10
system,
OpenVMS
8.4 BL2

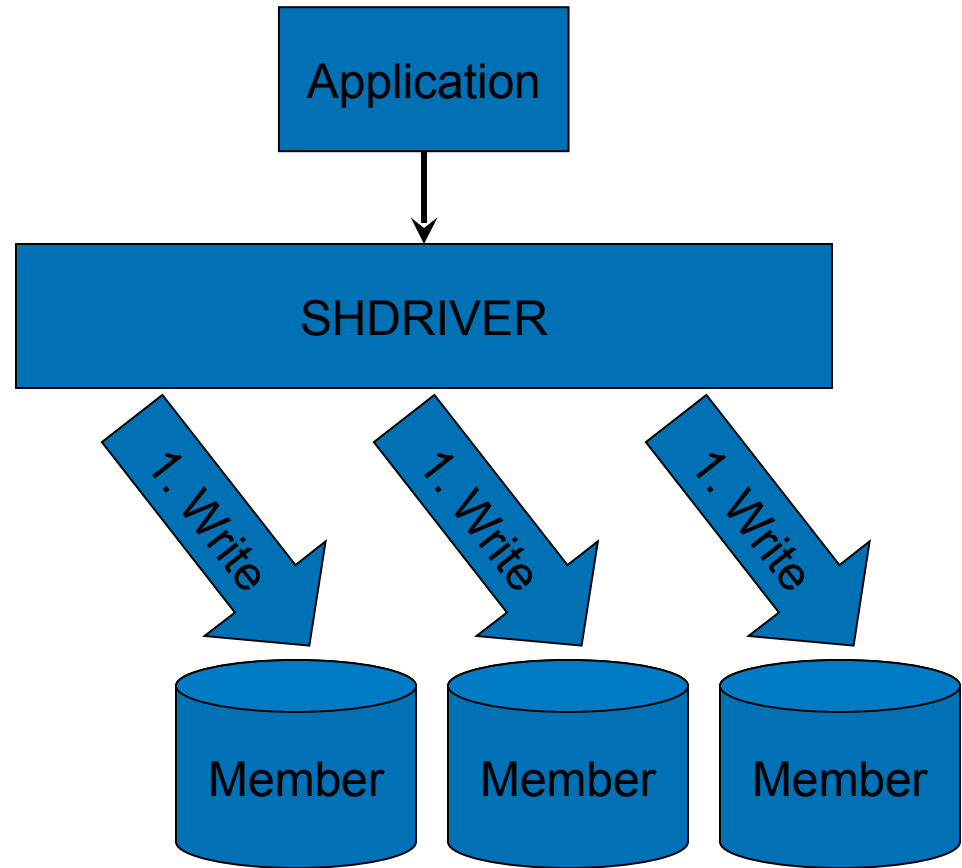


Performance Considerations

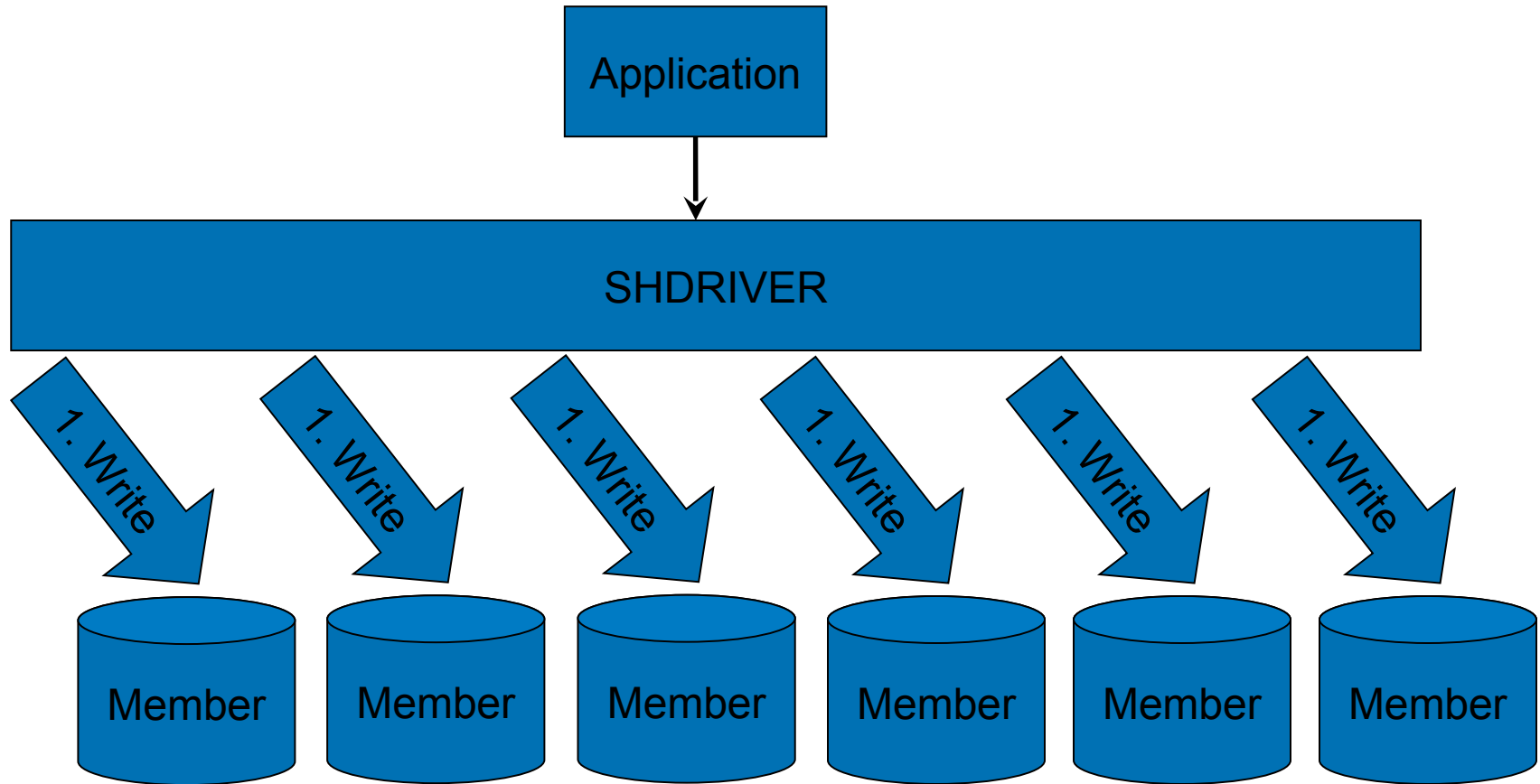
- Writes
 - Potential decrease in performance for writes
 - HBVS does Synchronous Mirroring:
 - Data is written to all disks (in parallel) for each write operation, and
 - Write is not completed until writes to all members are completed
 - Slowest write operation of the 6 thus gates performance
 - Write-back cache in controllers may tend to “hide” much or all of this impact

Application Write during Steady-State

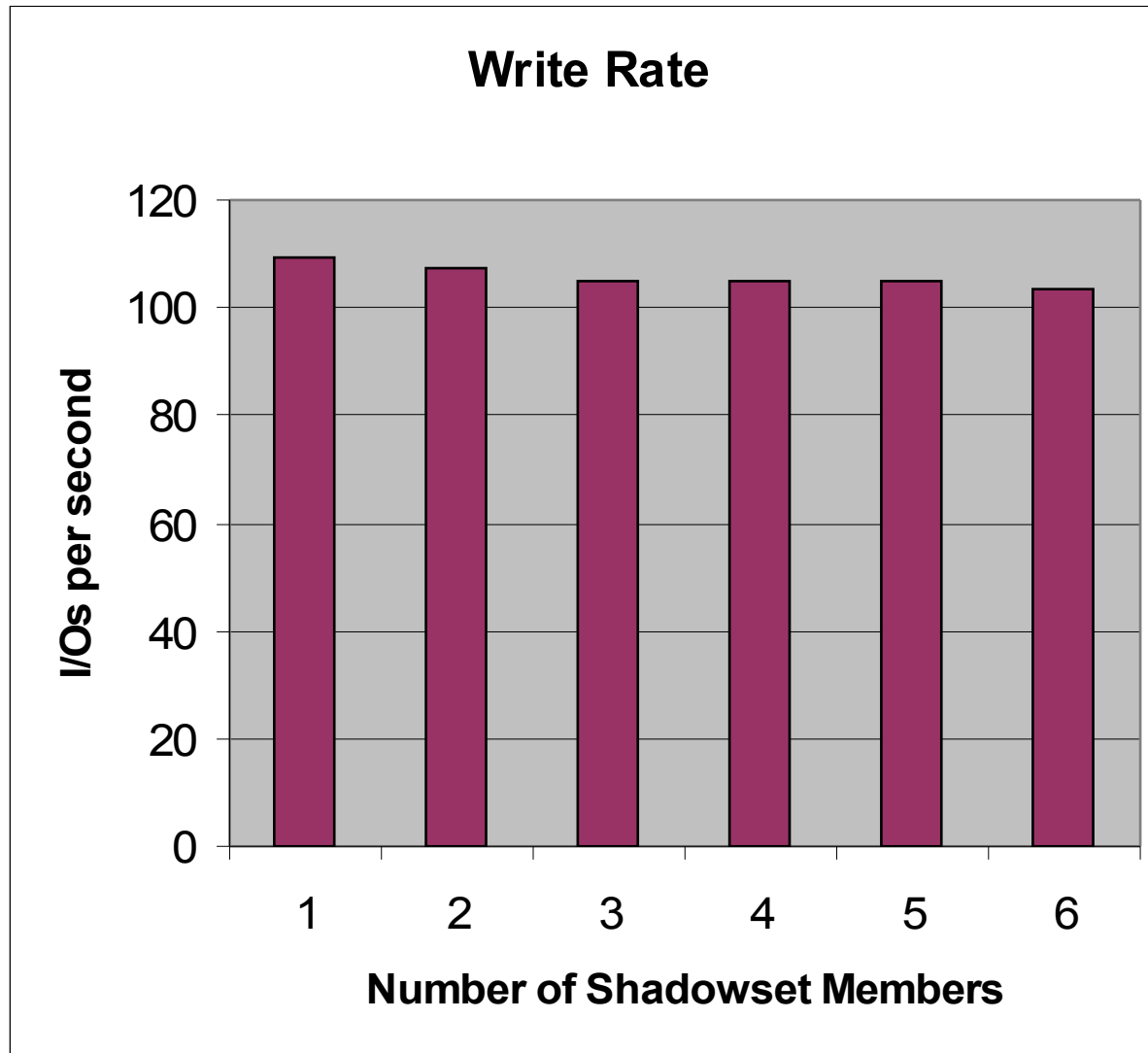
- Writes to all members in parallel



Application Write during Steady-State



Write Performance with 1-6 Members



Random 1-block I/Os, older local SCSI disks (thus no write-back cache), Alphaserver DS10 system, OpenVMS 8.4 BL2

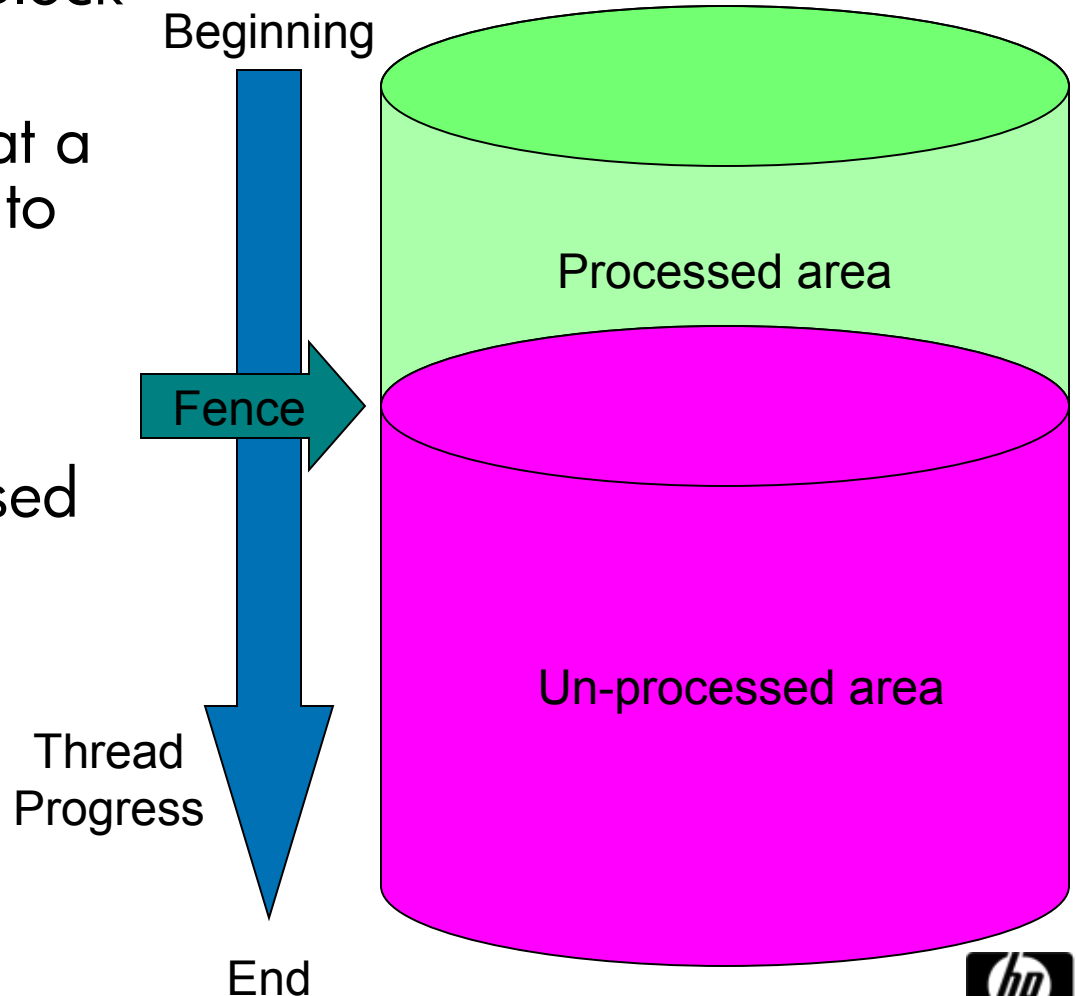


Performance Considerations

- Merges
 - Potential decrease in performance for reads ahead of the merge fence
 - Enabling *Mini-Merge* helps minimize the impact of this

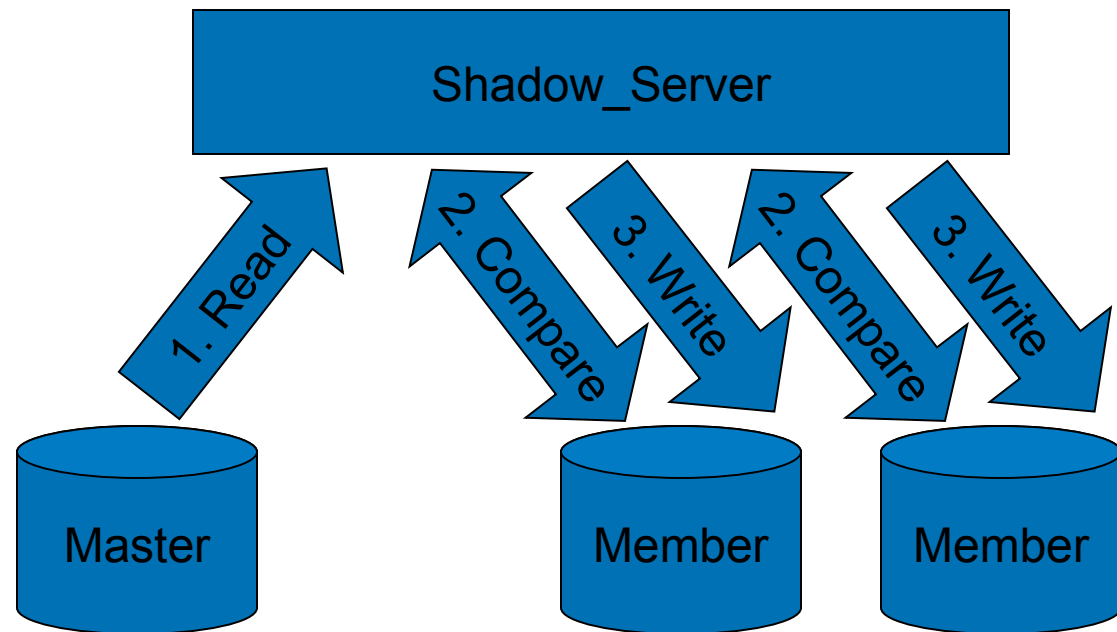
Full-Merge Thread Algorithm

- Start at first Logical Block on disk (LBN zero)
- Process 127 blocks at a time from beginning to end
- Symbolic "Fence" separates processed area from un-processed area

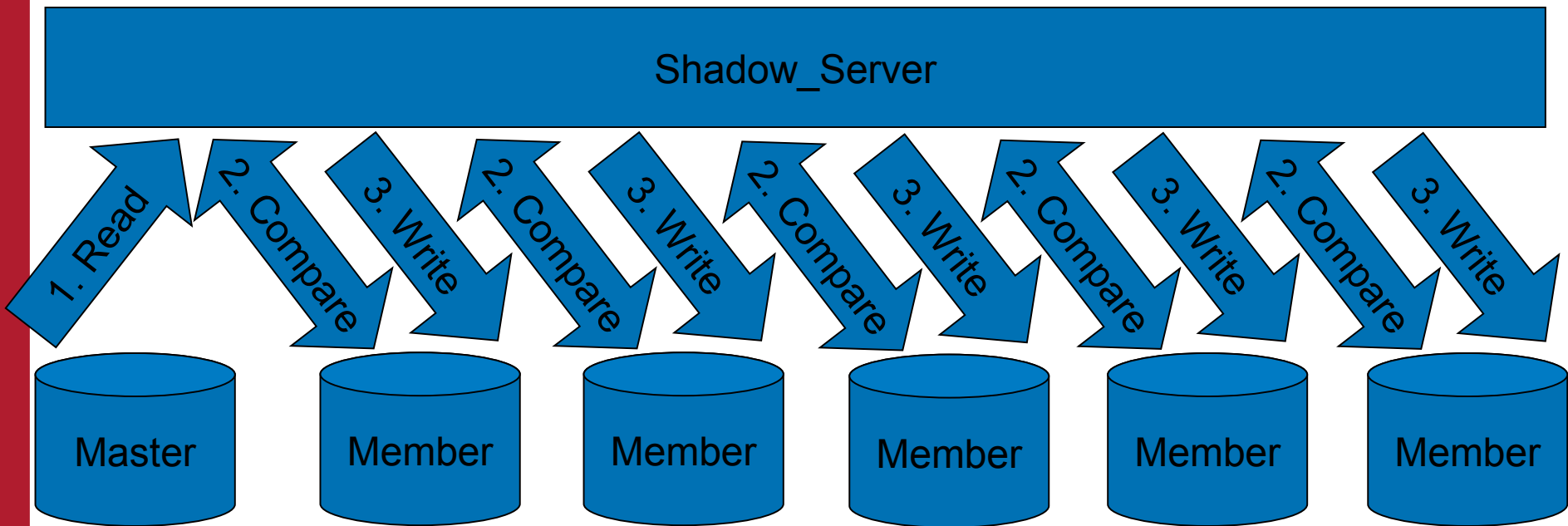


Merge Thread Algorithm

1. Read from master
2. Compare with other member(s)
3. If different, write data to target(s) and start over at Step 1.



Merge Thread Algorithm



Questions?



Speaker Contact Info



E-mail:

Keith.Parris@hp.com

Website:

<http://www2.openvms.org/kparris/>

Produced in cooperation with:

