

z/OS UNIX Shared File System Configuration - Overview and New Features

Ann Totten IBM Corporation

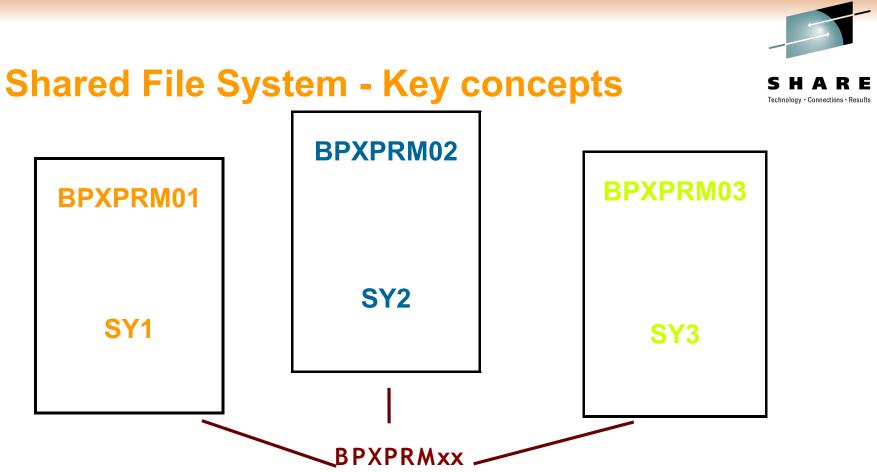
Wednesday, March 2, 2011: 4:30 PM-5:30 PM Session Number 9041





Session Topics

- Overview of z/OS UNIX shared file system support.
 - Setup and benefits.
 - File system configuration and multiple releases in one sharing group.
- File system ownership, management, including how to fail-safe the sysplex root file system.
- Actions required to migrate to z/OS Release 11 and 12.
- zFS shared file system changes in z/OS Release 11 and 12.
- Terminology on zFS sysplex-aware filesystems SHARE
- •² System and shell commands.

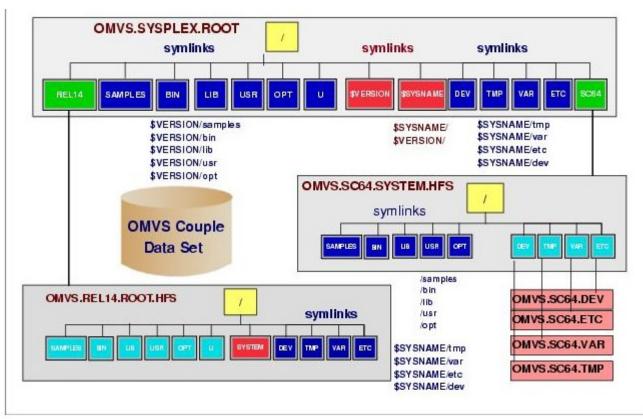


- All systems may share a common BPXPRMxx member that contains a ROOT statement and MOUNT statements. This is done through the use of system symbolics.
- ➡ The common BPXPRMxx member also contains the SYSPLEX(YES) statement.
- → Each system may specify a BPXPRMxx member that contains system specific limits.
- → All systems should have the same PFSs and any necessary subsystems required by the PFS started. For example, NFS client PFS requires the TCP/IP subsystem be started and a network connection configured.



Hierarchical file system concepts

 Figure 6-27 All the z/OS UNIX file sharing structures used in a sysplex sharing environment. Source: Redbook: UNIX System Services z/OS Version 1 Release 7 Implementation (ISBN 073849609X -IBM Form Number SG24-7035-01)





File system structures in a shared file system configuration



Sysplex root file system

- The sysplex root is a file system that is used as the sysplex-wide root. Only one sysplex root is allowed for all systems participating in a shared FS environment.
- System-specific file system
 - Directories in the system-specific file system data set are used as mount points, specifically for /etc, /var, /tmp, and /dev.
- Version file system
 - You can use one version file system for each set of systems participating in shared FS and that are at the same release level.





Shared File System - Key concepts

- All participating systems communicate with each other using coupling facility or channel-to-channel connections.
- BPXMCDS couple data set is a complex record of participating systems and the file systems in use throughout the plex.

SY1 RECORD

SY2 RECORD SY3 RECORD

SY4 RECORD

OMVS Couple Data Set System Record

 Contains information about each system in the shared FS group.

Changes as systems join or leave the sysplex.

How to see the system information in CDS:



MODIFY BPXOINIT, FILESYS=DISPLAY

BPXF242I 2011/02/08 08.34.28 MODIFY BPX0INIT, FILESYS=DISPLAY, GLOBAL

- SYSTEM LFS VERSION ---STATUS----- RECOMMENDED ACTION
- NP5 1. 12. 0 VERIFIED NONE
- NP6 1. 12. 0 VERIFIED NONE
- NP7 1. 12. 0 VERIFIED NONE
- CDS VERSION= 2 MIN LFS VERSION= 1. 12. 0
- DEVICE NUMBER OF LAST MOUNT= 372
- MAXIMUM MOUNT ENTRIES = 4000 MOUNT ENTRIES IN USE = 356
- MAXIMUM AMTRULES = 300 AMTRULES IN USE = 8
- MAXSYSTEM= 12

ALTROOT= USSZFS.ALTROOT.ZFS

SYSTEMS PERFORMING UNMOUNT

(Since 2011/02/04 00.35.11)

NUMBER OF UNMOUNTS IN PROGRESS= 1

NP6

ACTIVE QUEUE



Shared File System - Key concepts, cont



rechnology · Connections · Resul

PLEX.SYSPLEX.ROOT

PLEX.V1R10.VERFS

PLEX.V1R12.VERFS

PLEX.SY1.FS

PLEX.SY2.FS

PLEX.SY3.FS

PLEX.SY2.FS

OMVS Couple Data Set Mount Record

Used to keep hierarchy consistent across the sysplex.

•Provides information about each file system in use throughout the shared file system group.

•Changes as file systems are mounted, unmounted, moved, recovered, remounted or become unowned.



MODIFY B	PXOINIT,	FILESYS=D	ISPLAY,	, ALL
----------	----------	-----------	---------	-------

BPXF0351	2011/02/08	08.52.12	MODIFY	BPXOINIT, FILESY	YS=DISPLAY
----------	------------	----------	--------	------------------	------------

NAME	DEVICE	MODE

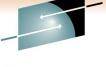
OMVSSPA.SVT.SYSPLEX.ZFS 1 RDWR

PATH=/

- LOCAL STATUS=ACTIVE STATUS=ACTIVE
- OWNER=NP4 **RECOVERY OWNER=NP4** AUTOMOVE=Y PFSMOVE=Y
- TYPENAME=ZFS MOUNTPOINT DEVICE= 0
- LFSFLAGS=08000000 ENTRY FLAGS=D0054000 FLAGS=40000200
- LOCAL FLAGS=40000200 LOCAL LFSFLAGS=28000000
- ACTIVECHK =0000000 LFSFLAGS2 =D8000000

...etc...





RE

SHARE Technology - Congections - Desuits

Shared file system terminology

- Non-Sysplex aware file system (sysplex-unaware)
 - The file system requires it to be accessed through the remote owning system from all other systems in the shared file system configuration. There is only one connection for update at a time for a particular mount mode (read-only or read-write).

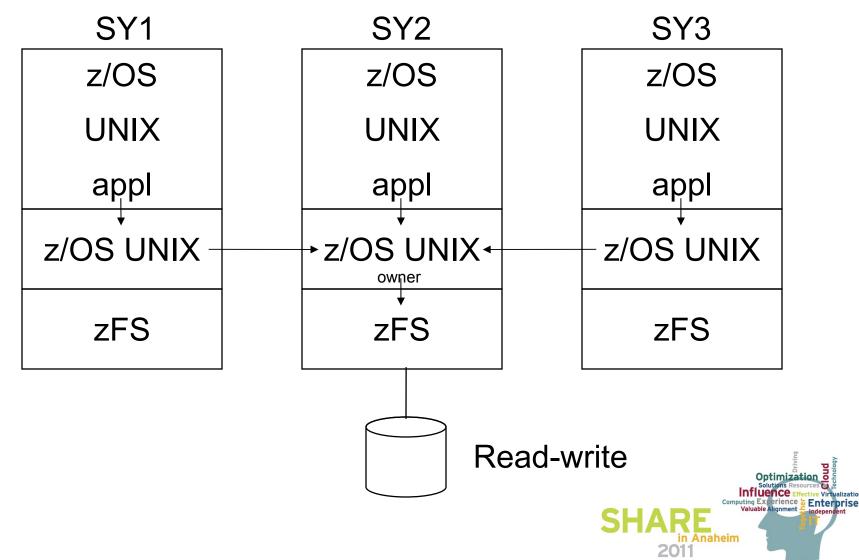
Function Shipping

- Function shipping means that a request is forwarded to the owning system and the response is returned back to the requestor through XCF communications.
- Sysplex-aware file system
 - The file system is locally mounted on every system and file requests are handled by the local PFS.



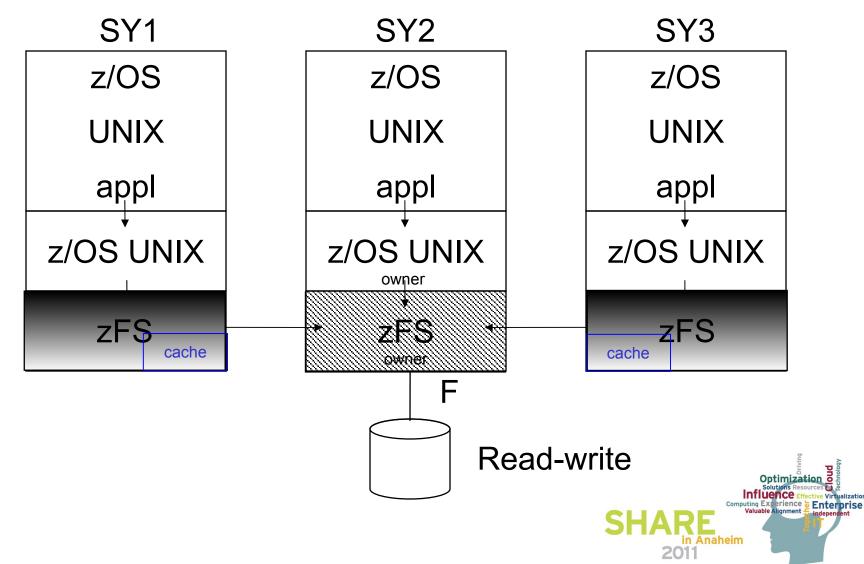
Non-sysplex aware file systems: z/OS UNIX does the function shipping of i/o requests





zFS shared file system, R/W sysplex-aware: zFS does the i/o (z/OS R11 and later)





File system ownership



z/OS UNIX owner

- System where the original mount is issued. Administrator can still control this via /usr/sbin/chmount or the other externals. Significant when the zFS R/W file system is mounted non sysplex-aware.
- Use the D OMVS, F or df -v or ishell to query the z/OS UNIX owner.

zFS owner

- System where the actual i/o is issued when the zFS R/W file system is sysplex-aware.
 - At mount time, this is the same system as the z/OS UNIX owner.
 - It is independent of the z/OS UNIX owner.
 - zFS will move the owner based on i/o activity.
- Use the zfsadm lsaggr command to query the zFS owner pr



File system movement to new owner

- Planned outages:
 - The modify omvs, shutdown command stops z/OS UNIX processes, synchronizes data to the file systems and unmounts or moves ownership in a shared file system environment. The shutdown succeeds only if all non-permanent z/OS UNIX processes end, all permanent processes are successfully checkpointed, and if all physical file systems are successfully quiesced.
 - New z/OS UNIX owner will follow the automove rules.
 - New zFS owner is determined by zFS and may or may not be the same as the z/OS UNIX owner.
 - The modify omvs, restart command restarts and reinitializes the z/OS UNIX environment.





File system movement to new owner

- Unplanned outages:
 - New z/OS UNIX owner will follow the automove rules.
 - New zFS owner is determined by zFS and may or may not be the same as the z/OS UNIX owner.





Fail-safe the sysplex root file system

- In a sysplex configuration, the alternate sysplex root file system is a hot standby for the sysplex root file system that is used to replace the current sysplex root file system when the sysplex root file system becomes unowned.
 - The alternate sysplex root file system is established by using the ALTROOT statement in the BPXPRMxx parmlib member during OMVS initialization or by using the SET OMVS command.



Steps for setting up the alternate sysplex root



- 1. Allocate a new file system to be used as the alternate sysplex root file system.
- 2. On the alternate sysplex root, set up the mount points and the symbolic links. The mount points and the symbolic links must be same as the ones on the current sysplex root.
- 3. Specify **ALTROOT** in the BPXPRMxx parmlib member with the mount point in the root directory of the current sysplex root file system.

Restriction: The ALTROOT mount point must not exceed 64 characters in length.

Example:

```
ALTROOT FILESYSTEM('USSZFS.ALTROOT.ZFS')
```

MOUNTPOINT ('/mnt')

- 4. Make sure that all systems in the shared file system environment have direct access to the new file system and can locally mount it.
- 5. Process the **ALTROOT** statement by using the SET OMVS command or by initializing the OMVS with the updated BPXPRMxx parmlib member.

Example:

SET OMVS=(xx)



Steps for removing the alternate sysplex root



1. In the BPXPRMxx parmlib member, replace the ALTROOT FILESYSTEM statement with the following statement:

ALTROOT NONE

Because the ALTROOT NONE and ALTROOT FILESYSTEM statements are mutually exclusive, only one can be specified in the BPXPRMxx parmlib member.

Note: If concatenating parmlib members result in multiple ALTROOT statements, then the first parmlib member specified on the OMVS= operator command that contains the ALTROOT statement will take effect.

2. Issue a SET OMVS operator command to process the ALTROOT NONE statement.

Example:

SET OMVS=(XX)



Steps for dynamically replacing the alternate sysplex root



- 1. To verify that the sysplex root is locally mounted on all systems, issue:
- Ro *all, D OMVS, F, NAME=root_file_system_name
- You should see **CLIENT=N** for each system.
- 2. Allocate a new file system to be used as the new sysplex root file system.

Rules: see notes

- 3. On the new sysplex root, set up the mount points and the symbolic links. The mount points and the symbolic links must be same as the ones on the current sysplex root.
- 4. On any system in the shared file system configuration, issue:
- F OMVS,NEWROOT=new.root.file.system.name,COND=<YES|NO|FORCE>
 - YES Proceed conditionally.
 - **NO** Proceed unconditionally.
 - **FORCE** This option allows user to replace a failing sysplex root with the user-specified new sysplex root.
- 5. Update the TYPE parameter and name of the sysplex root file system in the BPXPRM*xx* member of SYS1.PARMLIB.





Migration to V1R11 / R12, New Terminology

- sysplex_admin_level (Interface level)
 - What it is:
 - zFS XCF communication level in shared file system environment
 - Where it is specified:
 - IOEPRMxx member sysplex_admin_level=1|2
 - How to query:
 - MODIFY ZFS, QUERY, LEVEL Last line will show:

```
sysplex(admin-only) interface(1)
```





Toleration with zFS R10, Step 1

- Install APAR OA25026 on all z/OS V1R10 systems. This is a conditioning function APAR for zFS on z/OS V1R11 and R12. Make APAR OA25026 active on all systems through a rolling IPL. You are now running with zFS sysplex_admin_level=1.
- Running with sysplex_admin_level=1 enables code that uses a simplification of the zFS XCF protocol between zFS members.
- Verify that you are running with the correct configuration by issuing the following on each image:

```
MODIFY ZFS, QUERY, LEVEL
```

```
IOEZ00639I zFS kernel: z/OS zSeries File System
Version 01.10.00 Service Level OAnnnnn - HZFS3A0.
Created on Tue Jan 6 20:02:25 EST 2009.
sysplex(admin-only) interface(1)
```





Toleration with zFS R10, Step 2

- After APAR OA25026 is active on each z/OS V1R10 system, specify the sysplex_admin_level=2 configuration option in the IOEPRMxx file(s). Make this level active on all z/OS V1R10 systems through another rolling IPL. This is the toleration function for zFS on z/OS V1R11 and R12. (The default for sysplex_admin_level is sysplex_admin_level=1).
- Verify that you are running with the correct configuration by issuing the following on each image:

MODIFY ZFS, QUERY, LEVEL

```
IOEZ00639I zFS kernel: z/OS zSeries File System
Version 01.10.00 Service Level OAnnnnn - HZFS3A0.
Created on Tue Jan 6 20:02:25 EST 2009.
sysplex(admin-only) interface(2)
```





Toleration with zFS R10, Notes

•sysplex_admin_level is defaulted to 1 in APAR OA25026

•You cannot skip step 1 and do only step 2 if you are running any z/OS V1R10 systems that do not have APAR OA25026 installed and active.

•Your systems must be running zFS V1R10 and above to support a zFS V1R11 or R12 environment. In other words, zFS V1R11 is not compatible with V1R8 and below.

•The *sysplex_admin_level* configuration option cannot be specified dynamically through the **zfsadm config** command.

•When all images are running at *sysplex_admin_level=2,* the XCF protocol used by zFS is simplified.

•The value must be 2 on all members of the sysplex in order to bring z/OS V1R11 or R12 zFS into the shared file system environment.

•In V1R11 and later, the *sysplex_admin_level* setting is ignored and set to 3.





New Terminology in R11/R12

sysplex=on|off

- What it is:
 - Specifies whether zFS should run sysplex-aware.
 - On specifies sysplex-aware for read-write file systems.
 - Off specifies non sysplex-aware for read-write file systems.
- Where it is specified:
 - IOEPRMxx member

sysplex=on

- Default: sysplex=off
- How to query:
 - MODIFY ZFS, QUERY, LEVEL
 - zfsadm configq -syslevel

Example output: sysplex(file) interface(3

Sysplex Aware Capability Expanded in R11/R12



sysplex-aware

- A zFS file system can be mounted sysplex-aware or non-sysplex aware. When a file system is mounted sysplex-aware, it means that the file system is locally mounted on every system. When a R/W file system is mounted non-sysplex aware, this is functionally equivalent to z/OS UNIX file system sharing.
- Term can be used to describe the way that zFS was started. If started with sysplex=filesys or sysplex=on, then zFS is said to be "sysplex-aware". That it, zFS is running capable of mounting file systems sysplex-aware.
- Unchanged:
 - Read-only mounted file systems are always sysplex-aware.
 - HFS mounted R/W in a shared file system environment are always non sysplex-aware.





Benefits of Sysplex-Aware zFS

- The zFS read-write file systems that are owned and accessed by any of the systems that are running zFS sysplex-aware typically see a performance benefit.
- The zFS sysplex sharing allow sysplex clients to cache data to reduce the number of calls sent to servers.
- zFS ownership move support allows ownership of aggregates to dynamically move to the highest usage system.





zFS sysplex-aware on a file system basis

- Enhanced function
- Provided via
 - zFS APAR OA29619 and
 - z/OS UNIX APAR OA29712
- Provides another option to implement zFS sysplex-aware functionality.
- Allows you to specify that individual zFS file systems should be sysplex-aware rather than all zFS file systems being sysplexaware.





New Terminology provided by APAR OA29619

sysplex=filesys

• What it is:

- Specifies that zFS should run sysplex-aware and support sysplex-aware behavior on a file system basis.
- Where it is specified:
 - IOEPRMxx member
 - sysplex=filesys
- How to query:
 - MODIFY ZFS, QUERY, LEVEL
 - zfsadm configq -syslevel
 - Example output: sysplex(filesys,rwshare) interface(3)
- IBM recommendation for R11 and later releases: sysplex=filesys





New Terminology provided by APAR OA29619

sysplex_filesys_sharemode

- What it is:
 - Specifies if zFS read-write file systems will be mounted sysplex-aware when a MOUNT is issued on that system without specifying either NORWSHARE or RWSHARE in the MOUNT PARM.
- Where it is specified:
 - IOEPRMxx member
 - sysplex_filesys_sharemode=rwshare
- Can be dynamically changed:
 - zfsadm config -sysplex_filesys_sharemode rwshare
- Default if not specified:
 - norwshare
- How to query:

MODIFY ZFS, QUERY, LEVEL

zfsadm configq -sysplex_filesys_sharemode





New Terminology provided by APAR OA29619

- Mount parameter, NORWSHARE or RWSHARE
 - What it is:
 - Specifies whether a zFS read-write mounted file system will be mounted sysplex-aware or non sysplex-aware.
 - Where it is specified:
 - On the mount statement. Example:

```
# /usr/sbin/mount -f USSZFS.WORK01.ZFS -t zfs -o 'RWSHARE' /zfs wkld/work1
# df -Pkv /zfs wkld/work1
Filesystem
                   1024-blocks
                                      Used Available Capacity Mounted on
                                     17281
                                                              2% /zfs wkld/work1
USSZFS.WORK01.ZFS
                       1693440
                                               1676159
ZFS, Read/Write, Device:1386, ACLS=Y
RWSHARE
File System Owner : NP4
                                                 Client=N
                                Automove=Y
Filetag : T=off
                  codeset=0
```





What else has changed in this output?



/usr/sbin/mount -f USSZFS.WORK01.ZFS -t zfs -o 'RWSHARE' -d NP5 /zfs_wkld/work1
df -Pkv /zfs_wkld/work1

Filesystem1024-blocksUsedAvailableCapacityMounted onUSSZFS.WORK01.ZFS16934401728116761592% /zfswkld/work1

ZFS, Read/Write, Device:1386, ACLS=Y

RWSHARE

File System Owner : NP5 Automove=Y Client=N

Filetag : T=off codeset=0

Aggregate Name : USSZFS.WORK01.ZFS

- What about the File System Owner?
 - Still the z/OS UNIX owner
- What about Automove setting?
 - Unchanged, z/OS UNIX still uses this when processing mount takeover
- What about Client=?
 - Yes, this is one way to tell if a system is accessing the file system directly.





How to determine if a R/W zFS is sysplex-aware?

zfsadm aggrinfo -long (shell command)

zfsadm aggrinfo -long -aggr USSZFS.WORK01.ZFS USSZFS.WORK01.ZFS (R/W COMP): 1676159 K free out of total 1693440

version 1.4

auditfid D6C5F8F2 F0F72168 0010

sysplex-aware

209519 free 8k blocks; 7 free 1K fragments

16952 K log file; 56 K filesystem table

248 K bitmap file





How to determine which system is the zFS owner?

zfsadm lsaggr (shell command)
zfsadm lsaggr | grep USSZFS.WORK01.ZFS
USSZFS.WORK01.ZFS

R/W

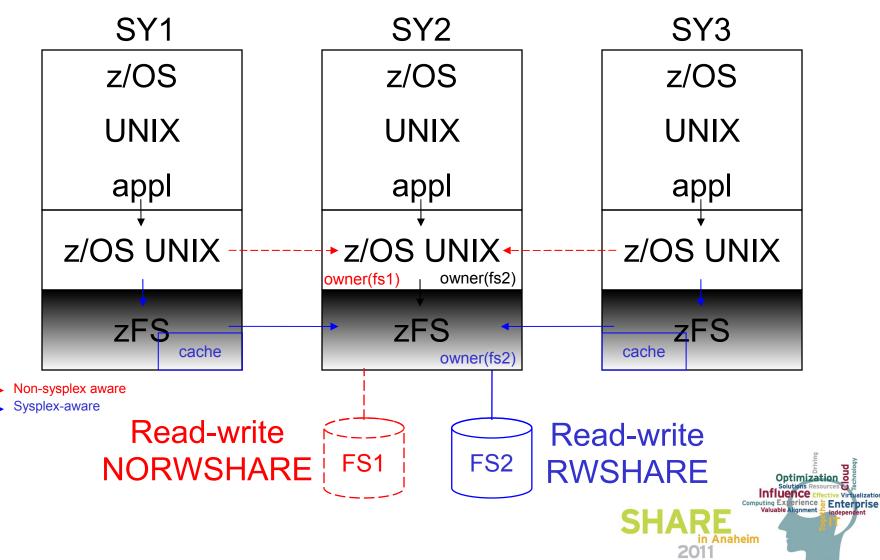
NP5





Technology • Connections • Results

zFS Release 11/12 sysplex=filesys support



S H A R E

zFS Release 11/12 sysplex=filesys support

- Recommendation is to run sysplex=filesys for greater flexibility in deciding which zFS file systems should be sysplex-aware.
- If you are running z/OS SMB server, the Fast Response Cache Accelerator support of the IBM HTTP Server for z/OS V5.3 or another server that uses the Register File Interest API – mount those file systems NORWSHARE.
- Restriction: While you can run some systems with zFS as sysplex-aware and some with zFS as non-sysplex aware, it is not recommended and will cause a reduction in performance.



zFS Release 11/12 sysplex=filesys support



What zFS file systems should be mounted sysplex-aware?

Typically, if you make a zFS read-write file system sysplex-aware, you see a performance improvement in most shared file system environments when accessing the data from a system that is not the zFS owner.

- R/W file systems where there are updates from more than one image.
- There is a tool available to aid in determining which zFS read-write file systems should be sysplex-aware. The tool and documentation are located at URL:
- ftp://ftp.software.ibm.com/s390/zos/tools/wjsfsmon
 It can also be found at the z/OS UNIX tools and Toys site
- http://www-03.ibm.com/systems/z/os/zos/features/unix/tools/





Health Checks

• ZOSMIGV1R11_ZFS_INTERFACELEVEL

• Description:

Verifies that the system is running sysplex_admin_level=2 for zFS V1R11 toleration support.

Reason for check:

zFS should be running at sysplex_admin_level=2 for all members of the sysplex. Once this is done, zFS V1R11 may be brought into the sysplex.

Applies to z/OS V1R9 and V1R10.





Health Checks

• USS_CLIENT_MOUNTS

• Description:

This check will generate an exception when a file system is found that is function shipping but could be mounted locally.

• Reason for check:

File systems should not function ship if they can be mounted locally to avoid performance degradation.

• Check applies to z/OS V1R10 and later.





Publications

- z/OS UNIX System Services Planning (GA22-7800) General Administration of z/OS UNIX file systems
- z/OS UNIX Command Reference (SA22-7802) confights command for HFS
- z/OS MVS System Messages Volume 9 (IGF-IWM) (SA22-7639) IGWxxxt messages for HFS
- z/OS UNIX System Services Messages and Codes (SA22-7807)
 z/OS UNIX return codes, z/OS UNIX reason codes, X'5Bxxrrrr' reason codes for HFS
- z/OS Distributed File Service zSeries File System Administration (SC24-5989) zFS Concepts and zfsadm command for zFS
 - http://www-03.ibm.com/systems/z/os/zos/bkserv/r11pdf/#dfs
- z/OS Distributed File Services Messages and Codes (SC24-5917) IOEZxxxt messages and X'EFxxrrrr' reason codes for zFS
- z/OS Distributed File Service zSeries File System Implementation (SG24-6580)
 - Redbook available (updated March 2009 to include z/OS V1R10)
 - http://www.redbooks.ibm.com/abstracts/sg246580.html?Open
- z/OS Version 1 Release 8 Implementation (SG24-7265)
 - Redbook available (contains zFS updates for z/OS V1R8)
 - http://www.redbooks.ibm.com/abstracts/sg247265.html?Open
- z/OS DFSMS[™] Access Method Services for Catalogs (SC26-7394) IDCAMS utility
- z/OS DFSMS[™] Storage Administration Reference (SC26-7402) ADRDSSU utility for backup





How to determine if a zFS is sysplex-aware?

MODIFY ZFS,QUERY,FILESET (system command)

IOEZ00438I Starting Query Command FILESETS.

File System Name	Aggr #	Flg	Operations	
USSZFS.WORK01.ZFS	38	AMS	2	

The Flg column indicates the aggregate status as attached (A) and either mounted (M), quiesced (Q), cloning (C), deleting a clone (D), or growing (G). If the aggregate is sysplex-aware for read-write, it will also indicate sysplex-aware (S).



SHARE Technology - Connections - Besults

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

IBM Language Environment z/OS

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

UNIX is a registered trademark of The Open Group in the United States and other countries.

SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to aon-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

