IBM Education Assistance for z/OS V2R2

Item: zFS 64bit
zFS performance query APIs
Element/Component: Distributed File Service zFS

Material current as of May 2015

Agenda

- Trademarks
- Presentation Objectives
- Overview
- Usage & Invocation
- Migration & Coexistence Considerations
- Installation
- Presentation Summary
- Appendix

Trademarks

- See url http://www.ibm.com/legal/copytrade.shtml for a list of trademarks.

Presentation Objectives

- zFS 64bit support
- zFS performance query APIs

Overview: zFS 64-bit support

- Problem Statement / Need Addressed
  - Limitation of 32 bit addressability
  - Metadata performance

- Solution
  - 64 bit addressability
  - New log method
  - Eliminate metadata backing cache and only use one metadata cache
  - Running zFS in OMVS address space

- Benefit / Value
  - Eliminates issues with running out of storage below the bar
  - Allows for much bigger caching and larger trace history
  - Improves metadata performance, especially for parallel updates to the same v5 directory.
  - Improves vnode operations

Usage & Invocation: zFS 64-bit support

- Log cache statistics in new format
  - **Statistics log cache information API (247)**
    - Version 1 returns structure API_LOG_STAT
    - Version 2 returns new structure API_NL_STATS
  - **zfsadm query -logcache** and **MODIFY ZFS,QUERY,LOG** support new stats.
- Transaction cache is removed
  - With improved logging method, it is no longer needed.
  - Details in Migration section
- Client cache is removed.
  - z/OS V1R2 cannot coexist with z/OS V2R2 so it is no longer needed.
  - Details in Migration section

Usage & Invocation: zFS 64-bit support

- Eliminate usage of metadata backing cache

- As 64 bit support allows zFS to obtain caches above the bar, no longer need to define a metaback cache in data spaces .
  - IOEFSPRM option metaback_cache_size is used for compatibilty.
  - zFS internally combines meta cache and metaback cache and allocates 1 cache in zFS address space storage.
  - IBM recommends to remove **metaback_cache_size** option from IOEFSPRM and add its value to **meta_cache_size** option.
  - Details in Migration section

Usage & Invocation: zFS 64-bit support

- Best practice Health Check on z/OS V2R2 and higher:
  - **ZFS_CACHE_REMOVALS**
  - Determines if running with user specified IOEFSPRM configuration options metaback_cache_size ,client_cache_size and tran_cache_size.
  - Specify any of the options will cause exception. IBM recommends not to specify these 3 options.
  - User override check parms:
    - Keywords: METABACK, CLIENT, TRANS
    - Values: ABSENCE or EXISTENCE
    - e.g. PARM('METABACK(EXISTENCE), CLIENT(EXISTENCE), TRANS(EXISTENCE)')
  - Active, low severity.

Usage & Invocation: zFS 64-bit support

- New **Statistics Above 2G Storage Information API** - STATOP_STORAGE_ABOVE (255) via API_STOR_STATS2
  - Support >2G storage report
  - MODIFY ZFS,QUERY,STORAGE,DETAILS provides heap free list for serviceability.

- **Statistics Storage Information API-** STATOP_STORAGE(241) uses API_STOR_STATS2 for Version 2

```
            zFS Primary Address Space >2G Stge Usage
            ----------------------------------------
Total Storage Above 2G Bar Available:          4294963200M
Total Storage Above 2G Bar Allocated:          955252736

Total Bytes Allocated by IOEFSCM (Stack+Heap): 3145728
IOEFSCM Heap Bytes Allocated:                  3145728
IOEFSCM Heap Pieces Allocated:             66
IOEFSCM Heap Allocation Requests:          66
IOEFSCM Heap Free Requests:                 0

Total Bytes Allocated by IOEFSKN (Stack+Heap): 935329792
Total Bytes Discarded (unbacked) by IOEFSKN:           0
IOEFSKN Heap Bytes Allocated:                  905722338
IOEFSKN Heap Pieces Allocated:         252905
IOEFSKN Heap Allocation Requests:      252951
IOEFSKN Heap Free Requests:                46
…
```

Usage & Invocation: zFS 64-bit support

- In V2R2, zFS can run in the OMVS address space : 10%-20% cpu reduction (workload dependent)

- Decide if want zFS running in OMVS
  - No ASNAME keyword in FILESYSTYPE statement in the BPXPRMxx parmlib member.
  - To specify zFS configuration parms:
    - IOEPRMxx parmlib
      or
    - IOEZPRM DD statement in OMVS proc
  - New MODIFY OMVS, pfs=zfs,*cmd*
    - available whether zFS is in its own address space or in the OMVS address space.
    - e.g. MODIFY OMVS,pfs=zfs,query,all
  - If OMVS does not use the value defined in IBM-supplied PPT (program property table) , ensure the OMVS id has the proper privilege as the zFS user id.
    -

Usage & Invocation: zFS 64-bit support

- With 64 bit support, some caches now support larger value range

IOEFSPRM config options	Old range	new range
vnode_cache_size	32 - 500,000	1000 - 10,000,000
meta_cache_size	1M – 1024M	1M – 64G
token_cache_size	20480 – 2,621,440	20480 – 20,000,000
trace_table_size	1M – 2048M	1M - 65535M
xcf_trace_table_size	1M – 2048M	1M - 65535M

- **Larger numbers use the following suffixes:**

t	units of 1,000.	for counters
m	units of 1,000,000.	
b	units of 1,000,000,000	
tr	units of 1,000,000,000,000	
K	units of 1,024.	for storage sizes
M	units of 1,048,576.	
G	units of 1,073,741,824	
T	units of 1,099,511,627,776	

Migration & Coexistence Considerations: zFS 64-bit support

- **Toleration APAR OA46026 must be installed and active on all z/OS V1R13 and z/OS V2R1 systems prior to introducing z/OS V2R2.**
  - New format of log cache statistics
    - Allows down level systems to recognize the new logging method and run the new log recovery and return Version 1 output (API_LOG_STAT are mostly 0s).
    - Apps use **STATOP_LOG_CACHE (247)** to request Version 1 output should be updated to use Version 2 output.
    - **zfsadm query -logcache** and **MODIFY ZFS,QUERY,LOG** return new stats.
- Use SMP/E FIXCAT for z/OS V2.2 coexistence verification.

Migration & Coexistence Considerations

- Removal of **transaction cache** and **client cache.**
  - If using **IOEFSPRM** config option **tran_cache_size** or **client_cache_size,** size is ignored.
  - If using **Statistics APIs**
    - **STATOP_USER_CACHE(242)** *(returns remote VM_STATS 0s respectively for version 1 request; No remote VM_STATs for version 2 request)*

      - Update Version 1 request to Version 2 for new output
    - **STATOP_TRAN_CACHE(250)** *(returns 0s for version 1 request and nothing for version 2 request)*
      - - Use STATOP_LOG_CACHE (247) with Version 2 request for new output
  - If using **Query Config Option tran_cache_size setting(208), client_cache_size setting(231) API** or **Set Config Option tran_cache_size(160), client_cache_size(230) API**
    - no effect
  - If using commands
    - **zfsadm config** or **configquery -tran_cache_size| -client_cache_size:** no effect
    - **zfsadm query -trancache** now displays 0s : consider to remove the cmd
    - **MODIFY ZFS,QUERY,LFS** report now removes transaction cache : be aware

Migration & Coexistence Considerations: zFS 64-bit support

- Removal of metadata backing cache data space
  - Metaback cache size is combined into meta cache.
  - IBM recommends to remove metaback_cache_size option from IOEFSPRM after combining the sizes.
  - For compatibility, metaback_cache_size will be added to meta_cahe_size to get the total size of the metadata cache.

Installation: zFS 64-bit support

- Install toleration APAR **OA46026** for the down-level systems
- zFS configuration options **tran_cache_size** and **client_ cache _size** are ignored.

Overview: zFS query APIs

- Problem Statement / Need Addressed
  - zFS statistics wrapped too often.
  - Too much overhead when calling query APIs.
  - zFS did not show all statistics related to RWSHARE aggregates.
  - Need more detailed statistics per file system.

- Solution
  - 4 byte counters (version 1) → 8 byte counters (version 2)
  - 3 new sysplex related APIs.
  - New FSINFO function to obtain detailed file system information.

- Benefit / Value
  - Allow for monitoring statistics over a much longer period of time
  - Improve performance.
  - FSINFO provides more detailed information for single/multiple file systems in a faster and more flexible manner, including sysplex-wide information.

Usage & Invocation: zFS query APIs

- Existing APIs supports 8-byte counters
- STATOP_LOCKING (240)
- STATOP_STORAGE (241)
- STATOP_USER_CACHE (242)
- STATOP_IOCOUNTS (243, aka STATOP_IOREPORT1)
- STATOP_IOBYAGGR (244, aka STATOP_IOREPORT2)
- STATOP_IOBYDASD (245, aka STATOP_IOREPORT3)
- STATOP_KNPFS (246)
- STATOP_META_CACHE (248)
- STATOP_VNODE_CACHE (251)
- Affected **zfsadm query** commands and **MODIFY QUERY** commands

Usage & Invocation: zFS query APIs

- New sysplex related APIs
  - Statistics Sysplex Client Operation Info - STATOP_CTKC (**253**)
  - Server Token management Info - STATOP_STKM (**252**)
  - Statistics Sysplex Owner Operation - STATOP_SVI (**254**)

- New zfsadm query options
  - **zfsadm query -ctkc**
  - **zfsadm query -stkm**
  - **zfsadm query -svi**

- Existing **MODIFY ZFS,QUERY,CTKC|STKM|SVI** commands now support 8-byte counters

Usage & Invocation: zFS query APIs

- FSINFO provides:
  - zfsadm command
  - List Detailed File system API- ZFSCALL_FSINFO (0x40000013)
  - MODIFY command
  - Always supports 8 byte counters
- Recommend use FSINFO over List Aggregate Status (135 or 140) or List File system status (142)

Usage & Invocation: zFS query APIs

- Syntax:

zfsadm fsinfo [-aggregate *name* | -path *path_name* | -all]

[-basic |-owner | -full |-reset]

[-select *criteria* | -exceptions]

[-sort *sort_name*][-level][-help]

- Options:
  - aggregate *name*

    Specifies the name of the aggregate to be displayed. The aggregate name is not case-sensitive and is translated to uppercase. To specify multiple aggregates with similar names, use an asterisk (*) at the beginning, at the end, or both at the beginning and the end of name as a wildcard. If -**aggregate** *name* is specified with wildcards, the default display mode is -**basic**. Otherwise, the default display is -**owner**.

Usage & Invocation: zFS query APIs

- -path *path_name*

  Specifies the path name of a file or directory that is contained in the file system for which information is to be displayed. The path name is case-sensitive and can start with or without a slash (/). The default information display will be as if -**owner** were specified.

- -all

  Displays information for all aggregates in the sysplex. It is the default when -aggregate and -**path** are not specified. The default information display will be as if -**owner** were specified.

- -basic

  Displays a line of basic file system information for each specified file system. This option is the default in the following situations:

  - The -all option is specified but -**full**, -**owner**, and -**reset** are not specified.
  - None of -**aggregate**, -**all**, -**path**, -**full**, -**owner**, and -**reset** options are specified.
  - The -**sort** and -**exceptions** options are specified and neither -**full** nor -**owner** is specified.
  - The -**aggregate** option is specified with one or more wildcards.

Usage & Invocation: zFS query APIs

- -owner

  Displays only information that is maintained by the system owning each specified file system. This option is the default when -**aggregate** without wildcards is specified.

- -full

  Displays information that is maintained by the system owning each specified file system. It also displays information that is locally maintained by each system in the sysplex that has each specified file system locally mounted.

- -reset

  Resets zFS statistics relating to each specified file system. This option requires system administrator authority.

Usage & Invocation: zFS query APIs

- -exceptions Displays information about aggregate that has any exceptional conditions listed in the table. This option cannot be specified with -**reset**, -**path**, and -**select**. Information will be displayed by default as if the -**basic** option were specified.

Exceptions	**Apply to aggregates that ....**
CE	Had XCF communication failures between clients systems and owning systems. This typically means that applications have gotten timeout errors.
DA	Are marked damaged by the zFS salvager.
DI	Are disabled for reading and writing.
GD	Are disabled for dynamic grow.
GF	Had failed dynamic grow attempts
IE	Had disk IO errors.
L	Have less than 1 MB of free space, which means that increased XCF traffic is required for writing files.
Q	Are currently quiesced.
SE	Have returned ENOSPC errors to applications.
V5D	Shows aggregates that are disabled for conversion to version 1.5

Usage & Invocation: zFS query APIs

- -select *criteria*

  Indicates that each specified file system that matches the criteria is to be displayed. Information is displayed by default as if the -**basic** option were specified. The information that is displayed can also be sorted by using the -sort option.

  To use this option, specify one or more select criterias from the next page. Multiple criterias are separated by commas, such as ' **-select** Q,DI,L'

  This option cannot be specified with -**exceptions**, -**reset**, and -**path**.

Usage & Invocation: zFS query APIs

Criteria	Show aggregates that ....
CE	Had XCF communication failures between clients systems and owning systems. It typically means that applications have gotten timeout errors.
DA	Are marked damaged by the zFS salvager.
DI	Are disabled for reading and writing.
GD	Are disabled for dynamic grow.
GF	Had failed dynamic grow attempts
GR	Are currently being grown.
IE	Have returned ENOSPC errors to applications.
L	Have less than 1 MB of free space, which means that increased XCF traffic is required for writing files.
NS	Are mounted NORWSHARE.
OV	Contain extended (v5) directories that are using overflow pages.
Q	Are currently quiesced.
RQ	Had application activity.
RO	Are mounted read-only.
RW	Are mounted read/write.
RS	Are mounted RWSHARE.
SE	Have returned ENOSPC errors to applications.
TH	Have sysplex thrashing objects in them.
V4	Shows aggregates that are version 1.4.
V5	Shows aggregates that are version 1.5.
V5D	Shows aggregates that are disabled for conversion to version 1.5
WR	Had application write activity.

Usage & Invocation: zFS query APIs

- -sort *sort_name*

  Specifies that the information displayed is to be sorted as specified by the value of sort name. The default is sort by name. This option cannot be specified with -**reset**.

sort_name	**Function**
Name	Sort by file system name, in ascending order. This sorting option is the default.
Requests	Sort by the number of external requests that are made to the file system by user applications, in descending order. The most actively requested file systems are listed first.
Response	Sort by response time of requests to the file system, in descending order. The slower responding file systems are listed first.

- -level

  Prints the level of the zfsadm command. Except for -**help**, all other valid options that are specified with -**level** are ignored.

- -help

  Prints the online help for this command. All other valid options that are specified with this option are ignored.

Usage & Invocation: zFS query APIs

- FSINFO examples

- To display basic file system information for zFS aggregate PLEX.DCEIMGNK.FSINFO:

zfsadm fsinfo -aggregate PLEX.DCEIMGNK.FSINFO -basic

```
PLEX.DCEIMGNK.FSINFO                        DCEIMGNJ  RW,RS,Q,GF,GD,L,SE
Legend: RW=Read-write, Q=Quiesced, GF=Grow failed, GD=Grow disabled
        L=Low on space, RS=Mounted RWSHARE, SE=Space errors reported
```

- To display the status of the file system owner using a wildcard:

zfsadm fsinfo -aggregate PLEX.DCEIMGNJ.FS*

```
PLEX.DCEIMGNJ.FS1                         DCEIMGNJ   RW,NS
PLEX.DCEIMGNJ.FS2                         DCEIMGNJ   RW,RS
PLEX.DCEIMGNJ.FS3                         DCEIMGNJ   RW,NS
Legend: RW=Read-write,NS=Mounted NORWSHARE,RS=Mounted RWSHARE
```

Usage & Invocation: zFS query APIs

- To display file system owner status for zFS aggregate PLEX.DCE1MGNK.FSINFO:

zfsadm fsinfo -aggregate PLEX.DCEIMGNK.FSINFO -owner

```
File System Name: PLEX.DCEIMGNK.FSINFO
*** owner information ***
Owner:              DCEIMGNJ     Converttov5:               ON,ENABLED
Size:               8640K        Free 8K Blocks:            1054
Free 1K Fragments:  7            Log File Size:             112K
Bitmap Size:        8K           Anode Table Size:          8K
File System Objects:6            Version:                   1.5
Overflow Pages:     0            Overflow HighWater:        0
Thrashing Objects:  0            Thrashing Resolution:      0
Token Revocations:  8            Revocation Wait Time:      0
Devno:              51           Space Monitoring:          0,0
Quiescing System:   DCEIMGNJ     Quiescing Job Name:        SUIMGNJ
Quiescor ASID:      x4c          File System Grow:          ON,0
Status:             RW,RS,Q,GF,GD,L,SE
Audit Fid:          00000000 00000000 0000

File System Creation Time: Nov 5 15:15:54 2013
Time of Ownership:         Nov 5 15:25:32 2013
Statistics Reset Time:     Nov 5 15:25:32 2013
Quiesce Time:              Nov 5 15:28:39 2013
Last Grow Time:            n/a
Connected Clients:         DCEIMGNK

Legend: RW=Read-write, Q=Quiesced, GF=Grow failed, AGGRGROW disabled
        L=Low on space, RS=Mounted RWSHARE, SE=Space errors reported
```

Usage & Invocation: zFS query APIs

- To display sysplex-wide file system information for zFS aggregate PLEX.DCE1MGNK.FSINFO:

zfsadm fsinfo -aggregate PLEX.DCEIMGNK.FSINFO -full

```
....Skipped owner information (same as the last slide)...
   *** local data from system DCEIMGNJ (owner: DCEIMGNJ) ***
   Vnodes:              1               LFS Held Vnodes:         1
   Open Objects:        0               Tokens:                  3
   User Cache 4K Pages: 5               Metadata Cache 8K Pages: 6
   Application Reads:   167840          Avg. Read Resp. Time:    0.059
   Application Writes:  23460           Avg. Writes Resp. Time:  0.682
   Read XCF Calls:      0               Avg. Rd XCF Resp. Time:  0.000
   Write XCF Calls:     0               Avg. Wr XCF Resp. Time:  0.000
   ENOSPC Errors:       0               Disk IO Errors:          0
   XCF Comm. Failures:  0               Cancelled Operations:    0

   DDNAME:              SYS00004
   Mount Time:          Nov  6 09:46:44 2013

VOLSER PAV    Reads      KBytes     Writes     KBytes     Waits    Average
______ ___ __________ __________ __________ __________ __________ _________
CFC001   1         12         88      25777     304164      18798 1.032
______ ___ __________ __________ __________ __________ __________ _________
TOTALS             12         88      25777     304164      18798 1.032

   *** local data from system DCEIMGNK (owner: DCEIMGNJ) ***
   Vnodes:              3               LFS Held Vnodes:         2
   ....
   DDNAME:              SYS00004
   Mount Time:          Nov  6 09:46:44 2013

VOLSER PAV    Reads      KBytes     Writes     KBytes     Waits    Average
______ ___ __________ __________ __________ __________ __________ _________
CFC001   1          6         44       7240      53764          6 0.513
______ ___ __________ __________ __________ __________ __________ _________
TOTALS              6         44       7240      53764          6 0.513
```

Usage & Invocation: zFS query APIs

- New ZFSCALL_FSINFO API (0x40000013)
- zFS PFSCTL API (BPX1PCT) code for FSINFO

```
BPX1PCT(
        "ZFS     ", /* File system type followed by 5 blanks */
        0x40000013, /* ZFSCALL_FSINFO – fsinfo operation */
        parmlen, /* Length of parameter buffer */
        parmbuf, /* Address of parameter buffer */
        &rv,        /* return value */
        &rc,        /* return code */
        &rsn)       /* reason code */
```

- 2 subcommands :
  - query file system info (153)
    - Requires minimum buffer size is 10K for single-aggregate query and 64K for multi-aggregate query.
  - reset file system stats (154)
    - Requires minimum buffer size is 10K

Usage & Invocation: zFS query APIs

- Syntax:

Modify *procname*,**fsinfo[,{***aggrname* **| all}**

[,{full | basic | owner | reset}

[,{select=*criteria* | **exceptions}] [,sort=***sort_name***]]]**

–

- Multiple selection criterias are separated by blanks.
- Parms are positional.

- Example:

-To display basic file system status for all zFS aggregates that are quiesced, damaged or disabled and also to sort aggregate names by response time:

modify zfs,fsinfo,all,basic,select=Q DA DI,sort=response

Migration & Coexistence Considerations: zFS query APIs

- **Toleration APAR OA46026 must be installed and active on all z/OS V1R13 and z/OS V2R1 systems prior to introducing z/OS V2R2.**
  - 4 byte counter (Version 1 output) → 8 byte counter (Version 2 output)
    - Allows down level systems to tolerate Version 2 output requests by returning Version 1 data.
  - FSINFO
    - Allows down level systems to handle the new function.
- Use SMP/E FIXCAT for z/OS coexistence verification.

Installation

- Install toleration APAR **OA46026** for the down-level systems

Miscellaneous

- Removal of two zFS Health Checks
  - ZOSMIGV1R13_ZFS_FILESYS
  - ZOSMIGREC_ZFS_RM_MULTIFS

Presentation Summary

- 64 bit addressability
- 3 new APIs for querying sysplex statistics
- APIs supports 8-byte counters
- New logging method
- New function FSINFO

Appendix

- Publication references
  - *z/OS Distributed File Service zSeries File System Administration* (SC24-6887)
  - *z/OS Distributed File Service Messages and Codes* (SC24-6885)
  - *IBM Health Checker for z/OS User's Guide* (SC23-6843)