Grover Davidson – grover@us.ibm.com
Senior Software Engineer, AIX Development Support
IBM

Enterprise2014

pIN541
Hidden Features and Functionality in AIX

Introduction

- AIX has many changes that are useful to many customers but are not advertised
- Some of them are very useful to almost everyone
- These are some of those changes that may be very useful for your systems.
- If you have more, please let us know so we can add them to the list!

VMM Fork Policy

- AIX uses a (Copy On Write) COW fork policy as of AIX 6.1
- Unix rules dictate that just after a fork, the child will look exactly like the parent except that there will only be the forking thread running
- This includes memory map from the parent
- Since most child processes exec to load a new program right away, this is a waste to build the complete memory maps and then destroy them
- For applications that fork but do not exec or they exec after running for a period of time, this can be less efficient
  - Accessing a page for read access causes a page fault in both the parent and child
  - Accessing a page for modification causes a second page fault
- Other option is to use older mechanism of Copy On Reference (COR) fork
  - When the child references the page it gets a private copy right away
  - Uses more physical memory because read only pages reside in multiple process address space

VMM Fork Policy

- Controlled with the restricted tunable vmm_fork_policy
- Changing this is a system wide change and affects all processes
- New AIX process environment space tunable to address this issue
  - VMM_CNTRL
    - VMM_CNTRL=vmm_fork_policy=COR; export VMM_CNTRL
    - VMM_CNTRL=vmm_fork_policy=COW; export VMM_CNTRL
    - Test Oracle listener process using COR if there are a lot of new connections
- In addition to VMM_CNTRL, the XCOFF executable files format now supports a new flag set with:
  - bforkpolicy set with ledit or compiler:
    - ldedit -bforkpolicy:cor a.out
    - xlc -o a.out -bforkpolicy:cor ...
  - Override the VMM_CNTRL environment flag or vmm_fork_polcy tunable
- Available starting **AIX 6.1 TL9 and 7.1 TL 3**

lsof

- IBM now has lsof Version 4.85 available on the IBM downloads site
- Details can be found here:
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power+Systems/page/How+to+install+lsof+%28LiSt+Open+Files%29+utility
- Actual page to find download at is here :
  - https://www14.software.ibm.com/webapp/iwm/web/reg/download.do?source
- Also shipped as part of the AIX Expansion Pack
- Used to list details about open files, who has them open, etc

mount remount option

- Mount command now allows a remount option to change certain options without actually unmounting and mounting the file system again
- Allows mount options for a file system to be changed without interrupting the applications using the file system
- Has options for JFS2, NFS
- See man page for mount for full details
- Enabling noatime using remount:
  - mount –o remount,noatime /my/filesystem
- Disabling noatime:
  - mount –o remount,atime /my/filesystem
- Also support changing rbr, rbw, minpout, maxpout and other options
- If conflicting options are listed then the last option is used

lkdev command

- Allows the administrator to ‘lock’ a device out from being changed
- Locking the device also prevents rmdev from being used on it
- Device must be unlocked before commands can be used on it
- To lock/unlock, use:
  - lkdev –l /dev/device –a [-c comments-without-spaces] # Locks device
  - lkdev –l /dev/device –d # Unlocks device
- To view devices that are locked, use:
  - lkdev
- To see the devices locked and comments with the locks, use:
  - odmget CuLk
  - Note: the k is lower case
  - If device is not listed, it is not locked

Changeable ODM attributes

- New ODM feature adds a '+' to the changeable column in the ODM:

```
$ lsattr -El hdisk0
PCM             PCM/friend/vscsi                 Path Control Module        False
algorithm       fail_over                        Algorithm                  True
hcheck_cmd      test_unit_rdy                    Health Check Command       True+
hcheck_interval 0                                Health Check Interval      True+
hcheck_mode     nonactive                        Health Check Mode          True+
max_transfer    0x40000                          Maximum TRANSFER Size      True
pvid            00f6fd46dac527a80000000000000000 Physical volume identifier False
queue_depth     3                                Queue DEPTH                True
reserve_policy  no_reserve                       Reserve Policy             True+
```

- The True+ indicator shows the attribute can be changed *and* the change can be made without unconfiguring the device
- Also requires the use of the '-U' chdev flag to acknowledge you are changing the device without unconfiguring it:
  - chdev –l hdisk0 –a hcheck_interval=10 –U
  - Will fail without the '-U' flag
- Not all devices have this indicator

rendev command

- Renames devices
  - rendev –l hdisk5 –n hdisk100
- Restrictions do apply just as they do to normal devices
  - New device name cannot exceed 15 characters
  - If the new name exists or the new *raw* name (begins with an ‘r’) exists in /dev, the rename will fail
  - Do not rename disks the belong to rootvg or are expected to be added to rootvg
  - **The system may fail to boot if you change rootvg disks**
- Some devices may have special purposes (/dev/tty for example)
  - Renaming these may cause serious issues that require a system reload to recover from
- Test renaming the devices on a non-production system
  - Include rebooting as part of the system
  - Databases should also be started after the rename operation
- **Create a backup using mksysb before renaming a device**

/dev/null options

- /dev/null is the bit bucket
- Anything written to this file is discarded
- May shell scripts and other commands redirect output to the file
- Since /dev/null is accessed down the JFS2 file system path it goes through the JFS2 write call to reach the device driver
- The JFS2 write system call updates the last modified time for the /dev/null inode which also generates a JFS2 log transaction that is written synchronously
- If there are enough threads writing heavily to /dev/null concurrently:
  - The log traffic becomes significant
  - The writes become blocked by the log traffic
  - System becomes idle
- Enable slower updates for /dev/null only with:
  - raso –p –o devnull_lazytime=1
  - Dynamic
  - Modification time of /dev/null is only accurate to seconds
  - In most cases this is not a real issue

Changes to /

- Searching for a file begins at / when a file is referenced by full pathname
- Filenames are located one component at a time
- For /a/b/c:
  - / is searched for a and if not found an error is returned
  - /a is searched for b and if not found an error is returned
  - /a/b is searched for c and if not found an error is returned
- To ensure directories are not deleted at the instant the search is occurring an inuse count is adjusted
- This also prevent a file system from being unmounted while a search is happening
- As a result, there is significant activity to update the in use count of inode 2 of the root filesystem (‘/’)
- If ‘/’ is gone, the system is finished
- Understanding this we can remove the code that updates the inuse count of ‘/’
- The access time is also not longer updated for ‘/’
- modification time is still updated

Changes to /

- With 2 applications stating files (this retrieves the metadata) in 2 file system using full pathnames (beginning with '/') there is a lot of activity on '/'
- The 2 applications interfered with each other and anything else searching from / (ps always references /proc for example)
- With this change in place, the contention on / is removed and shifted one level down
- Can still have contention further down but does not impact searching from / down another path
- Available with:

```
IV44289 U858978   7100-03 bos.mp64 7.1.3.15 STATX/LOOKUPS/FILE OPENS
IV44289 U854697   7100-03 bos.mp64 7.1.3.0 STATX/LOOKUPS/FILE OPENS
IV44518 U859554   7100-02-04-1341 bos.mp64 7.1.2.17 STATX/LOOKUPS/FILE OPENS
IV47062 U860252   7100-01-09-1341 bos.mp64 7.1.1.20 STATX/LOOKUPS/FILE OPENS
IV44384 U859304   6100-09 bos.mp64 6.1.9.15 STATX/LOOKUPS/FILE OPENS
IV44384 U854672   6100-09 bos.mp64 6.1.9.0 STATX/LOOKUPS/FILE OPENS
IV47061 U862133    bos.mp64 6.1.8.18 STATX/LOOKUPS/FILE OPENS
IV43089 U861576    bos.mp64 6.1.7.21 STATX/LOOKUPS/FILE OPENS
```

- No additional changes are needed

Dynamic Partition Optimizer

- Use to enhance the affinity of cpu and memory.
- Tries to optimize the configuration so that cpu and memory are not distant
- Use lssrad –av to get current configuration
- See the following Virtualization Best Practices guide at the following website http://www14.software.ibm.com/webapp/set2/sas/f/best/home.html
- DPO related commands are run from the HMC.
- **Need to be at the following levels to get the following apars with 770 version of firmware:**
  6100-07-04 / 6100-08-03
  7100-00-04 / 7100-01-04 / 7100-02-03 / 7100-03-00

- These levels include the following fixes:
  - DPO NOT OPTIMIZING MEMORY DURING RUNTIME
  - SYSTEM CRASH DURING DPO
  - SYSTEM CRASH FOLLOWING DPO EVENT.
  - TOPOLOGY_UPDATE BIT NOT SET FOR DPO OPERATION
  - TRAP IN WAITPROC_FIND_RUN_QUEUE AFTER PMIG/PHIB/DPO

Please fill out an evaluation

- ibmtechu.com