|  |
| --- |
| General recommendations for AIX: |
|  |
|  |

Contents

Storage Array general and Queue Depth recommendations for AIX: 3

AIX recommendations: 6

HACMP recommendations: 9

Boot from SAN recommendations: 10

VIO (Virtual I/O/Advanced Power Virtualization) Considerations: 11

Multipathing and Disk Considerations: 12

# Storage Array general recommendations for AIX:

1. Install the ODM update provided by Hitachi Data Systems. This will set the default attributes, rw\_timeout=60 and queue\_depth=2. The Hitachi ODM update is a set of predefined attributes to identify disk devices. It is not a driver. It is required for support from IBM under the Service Implementation Plan with Hitachi Data Systems. Installing any of the Hitachi ODM updates requires a reboot. It is highly recommended to install the ODM updates (either the HDLM updates **or** the MPIO updates) at the same time so only one reboot is required.

Current ODM updates are:

HDLM/DMP/PowerPath:

* + - 1. - base
      2. Or 5.0.0.5 - update for device support

5.0.52.1 – update for AIX 5.2 includes dynamic tracking support

5.0.52.2 – update for HP co-existence

5.0.52.3 – update for HDME volumes

MPIO:

5.4.0.0 – base

5.4.0.1 - update for health check feature

5.4.0.2 – update for USPV

5.4.0.3 – update for the VSP

5.4.0.4 – update for timeout\_policy

5.4.0.5 -- update for HUSVM

5.4.1.0 – update for G1000/F1000

5.4.1.1 – update for Gx00

6.0.0.0 – base for the AMS2000

6.0.0.1 – update for HUS

Hitachi ODM updates detailed descriptions and download available at:

<https://knowledge.hds.com/HDS_Information/TUF/Servers/AIX_ODM_Updates>

1. Queue\_depth is set at the device level in AIX. The queue\_depth can be adjusted by issuing the “chdev –l hdiskx –a queue\_depth=x”. The device must be offline to make the change. The “-P” option can be used to update a device that is online and the change will take affect at the next reboot.

An initial queue\_depth setting of 8 is recommended. The maximum queue\_depth that can be specified is 32. The formula for calculating queue\_depth is the number of commands per port divided by the number of devices (hdisk). When using host storage domains, the queued commands are shared. The commands per port is by physical port.

The following calculations are used for each subsystem:

\_Each **USPV/USPVM** port can queue up to **2048** commands.

\_Each **VSP/HUSVM** port can queue up to **2048** commands.

\_Each **VSP G Series** port can queue **2048** commands (rule of thumb).

\_Each **Gxx/G1000** port can queue **2048** commands

\_Each **HUS/AMS** port can queue up to **512** commands.

Queue Depth ODM default can be set with CHDEF (MPIO environment) or DLMCHPDATR (HDLM environment) commands:

***dlmchpdattr*** *–a queue\_depth=32 (HDLM);* ***chdef*** *–a reserve\_policy=no\_reserve –c disk –s fcp –t htcvspmpio (MPIO)*

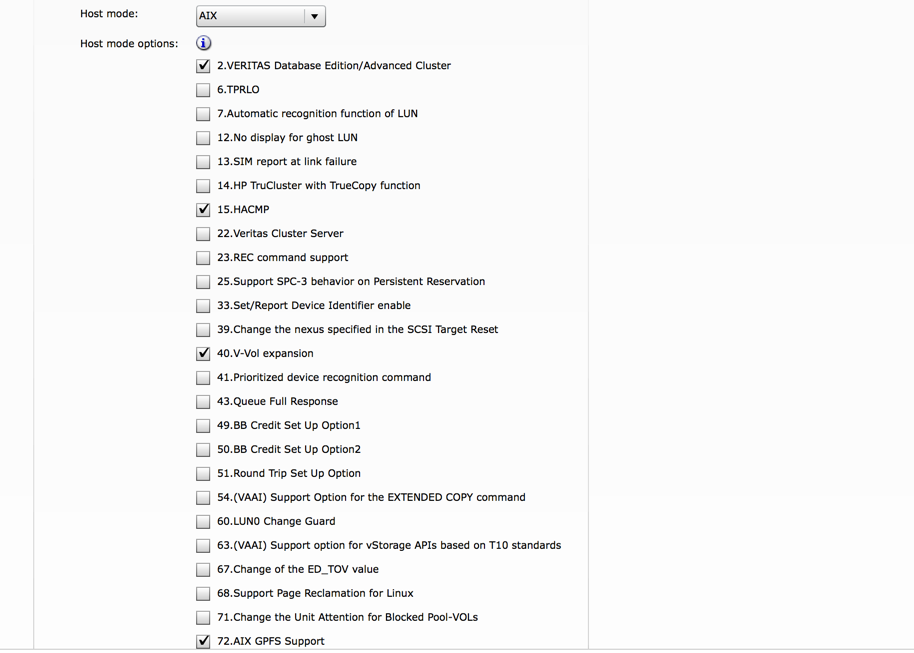
1. For optimum performance, it is not recommended that servers share a port. If multiple hosts do share a port, consider using Priority Port Control to limit the workload of the non-critical servers. When sharing ports, please note the following:

All HBA’s should be of the same type and firmware level on HDLM controlled paths on a host.  
  
No other path management type software should share the **same** HBAs as HDLM.   
  
Zoning should be performed at the HBA level.

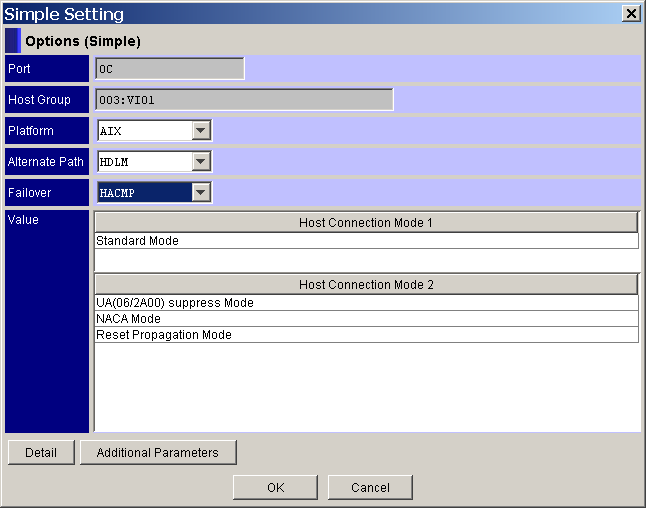
Multiple hosts accessing the same port are not required to have the same firmware/driver levels as long as the versions are supported.

1. For all subsystems, ensure that the “AIX” option is selected.
2. Also, host mode options “**2- Veritas Database Edition**” should be set for **HDLM**. “15 - HACMP” should be set if using PowerHA. Also, set “40 – Vvol expansion” is using HDP/HDT. For GPFS PR\_shared support, use HMO 72.
3. Host mode **option 2 with HDLM** is critical for **Live Partition Mobility** (LPM), High Availability Manager and GPFS.

Enterprise:

****

For Midrange subsystems port options, choose platform AIX, alternate path HDLM, NACA enabled, and failover HACMP, if appropriate:



NOTE: On the AMS2000 subsystems, setting AIX will set the proper options automatically.

HDLM requirements on Midrange with Persistent Reserves:

|  |  |
| --- | --- |
| **8** | Set the Host Mode Option [Unique Reserve Mode 1] ON. |
| **9** | When using "Simple Setteing" in Edit Host Group window:  Select VCS from the Middleware pulldown. When using "Additional setteing" in Edit Host Group window:  Select the "Unique Reseve Mode 1" checkbox. |

# AIX recommendations:

1. Always consult Hitachi Data Systems for interoperability of San components. Usually if a particular configuration is not supported, it is because there are known issues. Include your Hitachi Data Systems representative as a key component of any new implementation or change in your environment.

<https://support.hds.com/en_us/interoperability.html>

1. Use the latest HBA microcode available. HBA microcode levels should be check periodically. Microcode can be downloaded from:

<http://www14.software.ibm.com/webapp/set2/firmware/gjsn>

1. Check for maintenance (ptfs) that affect the driver. Occasionally there are ptfs that affect performance. It is always recommended by IBM and Hitachi Data Systems to be at the latest level. Check technology levels at:

<http://www-933.ibm.com/eserver/support/fixes/fixcentral>

1. For Oracle with filesystems, always use “aio” asynchronous I/O. This should be enabled and the min and max servers reviewed. For more information:

<http://publibn.boulder.ibm.com/doc_link/en_US/a_doc_lib/aixbman/prftungd/2365c812.htm>

<http://download.oracle.com/docs/cd/B28359_01/server.111/b32009/appa_aix.htm>

Oracle Architecture and Tuning on AIX:

<http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP100883>

1. Use raw logical volumes for databases or use JFS2 filesystems with the cio option. The cio mount option can give raw logical volume speed with the benefits of using a JFS2 filesystems at AIX 5.2 ML01. Do not use –cio for filesystems containing Oracle libraries or executables. The CIO option is only intended for applications that implement their own serialization like Oracle. Always check with your database/application vendor prior to using CIO.

<http://www-03.ibm.com/systems/p/os/aix/whitepapers/db_perf_aix.pdf>

1. Spread the logical volumes across all the disks in a volume group. This is an option when creating logical volumes (the –e option or maximum range in smit). Understand your logical to physical disk layout. Hdisks can be in the same physical disk group. Use lvmstat to ensure the workload is balanced across logical volume partitions:

<http://publib.boulder.ibm.com/infocenter/systems/index.jsp?topic=/com.ibm.aix.prftungd/doc/prftungd/lvm_perf_mon_lvmstat.htm&tocNode=int_13139>

1. Use vmtune (or vmo/ioo) to review JFS/JFS2 tuning parameters (fsbufwaitcnt and psbufwaitcnt). Ensure that the JFS/JFS2 buffers are not underallocated. Also check LVM and VMM counts (minpgahead/maxpgahead, numfsbufs, lvm\_bufcnt, hd\_pbuf\_cnt)

<http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.cmds/doc/aixcmds6/vmo.htm>

<http://publib.boulder.ibm.com/infocenter/pseries/v5r3/index.jsp?topic=/com.ibm.aix.cmds/doc/aixcmds3/ioo.htm>

AIX 6.1 has changed the default settings for VMM (AIX 6.1 defaults also apply to AIX 7.1):



1. Monitor on a regular basis using iostat (new options –ts –ta), vmstat, sar. Use filemon if you suspect a problem, however, remember that filemon does start a trace and can impact system performance.

<http://publibn.boulder.ibm.com/doc_link/en_US/a_doc_lib/cmds/aixcmds2/filemon.htm>

New iostat option that will help monitor queuing “iostat –D”. This example uses the –a option to monitor the adapter as well

# iostat –aD –d hdisk9 10

Adapter:

fcs1 xfer: bps tps bread bwrtn

1.6M 20.2 0.0 1.6M

Disks:

hdisk9 xfer: %tm\_act bps tps bread bwrtn

2.2 1.6M 20.0 0.0 1.6M

read: rps avgserv minserv maxserv timeouts fails

0.0 0.0 0.0 0.0 0 0

write: wps avgserv minserv maxserv timeouts fails

20.0 4.4 0.9 8.9 0 0

**queue:** avgtime mintime maxtime avgwqsz avgsqsz **sqfull**

125.5 0.0 121.1 0.7 0.0 **39.6**

| **Wait Queue Service Metrics (queue):** | **Restriction:**  **These metrics are not applicable for tapes.** |
| --- | --- |
| avgtime | Indicates the average time spent by a transfer request in the wait queue. Different suffixes are used to represent the unit of time. Default is in milliseconds. |
| mintime | Indicates the minimum time spent by a transfer request in the wait queue. Different suffixes are used to represent the unit of time. Default is in milliseconds. |
| maxtime | Indicates the maximum time spent by a transfer request in the wait queue. Different suffixes are used to represent the unit of time. Default is in milliseconds. |
| avgwqsz | Indicates the average wait queue size. |
| avgsqsz | Indicates the average service queue size. |
| sqfull | Indicates the number of times the service queue becomes full (that is, the disk is not accepting any more service requests) per second. |

Another new command in AIX 5.3 is fcstat which displays statistics of the fibre channel driver:

<http://publib.boulder.ibm.com/infocenter/systems/index.jsp?topic=/com.ibm.aix.cmds/doc/aixcmds2/fcstat.htm>

1. Consider adding more adapters/storage ports to increase throughput. Also, consider multi-pathing for greater availability. AIX native multipathing, MPIO, is supported on Enterprise subsystems (G1000/F1000, Gxx, HUS-VM) and the AMS2000/HUS. .
2. The default max\_transfer of 256KB is specified on the hdisks. This should be adequate for most workloads but may need to be increased for large sequential I/O’s. The LTG size for volume groups may also need to be checked in AIX 5.3. If the LTG size is larger than the max\_transfer size then there may be problems varying on or extending the volume group.
3. When using a switch change the fibre channel adapter to enable fast fail and dynamic tracking:

# chdev –l fscsix –a fc\_err\_recov=fast\_fail –a dyntrk=yes

1. Review the error report (errpt) on a regular basis and investigate any disk, I/O, or LVM errors.
2. Review the hba settings for num\_cmd\_elems and change. Monitor using the fcstat command.

**IBM MPIO Best Practices for AIX:**

<https://www.ibm.com/developerworks/aix/library/au-aix-mpio/>

# 

# HACMP recommendations:

1. Use enhanced concurrent volume groups to enable fast disk takeover.
2. When using HDLM, always define the custom disk method.
3. Ensure the appropriate settings are used on the subsystem (see Subsystem notes above).
4. Multi-pathing (either MPIO or HDLM) is required with HACMP to eliminate the hba or path as a single point of failure.
5. Heartbeat over disk is supported.
6. Truecopy/HUR can be implemented as a service with HACMP. PowerHA SystemMirror 6.1 Enterprise Edition also supports Truecopy/HUR and is integrated into PowerHA.
7. When using MPIO (default PCM), the SCSI-2 reserves must be disabled and the algorithm can be changed to round robin:

# chdev –l hdiskx –a reserve\_policy=no\_reserve –a algorithm=round\_robin

# Boot from SAN recommendations:

1. Always review the interoperability matrix to ensure the combination of hba/switch/storage subsystem is supported. Check the latest hba microcode to ensure booting from the SAN is enabled – this is not a problem on newer adapters.
2. Review and follow the OS vendors recommendations for booting from the SAN.
3. Certain HDS maintenance operations may have an affect on the Host System I/O to the SAN Boot device. HDS provided procedures should be followed.
4. Multi-pathing is recommended when booting from SAN either with HDLM or MPIO.
   1. HDLM 5.9 supports SAN boot
   2. HDLM 5.8 does not support SAN boot. The boot device(s) must be excluded from HDLM’s control, the use of two hdisks and LVM mirroring is recommended.
5. Allocating the OS (San Boot) disk and heavily used data disks in the same RAID groups should be avoided since the Data disk’s heavy i/o could impact the OS system I/O access to the system data. The OS (San Boot) disk and Data Disks should be in different RAID groups. Do not install many OS system disks in one RAID group (Four or less are recommended).
6. It is possible to change a boot disk definition to remove or change the ODM update, for example, migrating from MPIO to HDLM. However, care must be taken to ensure the disk is still bootable. Please contact HDS for specific instructions on changing boot disk definitions. Consider using virtual scsi devices for boot to eliminate this.
7. Set the reserve\_policy to no\_reserve for San boot devices.

# VIO (Virtual I/O/Advanced Power Virtualization) Considerations:

1. Virtual SCSI
   1. HDLM or MPIO is used/installed in the Virtual I/O servers
      1. The appropriate ODM must be installed
         1. 5.0.0.1 base and updates for HDLM
         2. 5.4.0.0 base and updates for MPIO
   2. The reserve policy **must** be changed to no\_reserve if using dual VIO servers.
      1. chdev –l hdiskx –a reserve\_policy=no\_reserve
      2. For HDLM, use the dlmchpattr command to permanently change attributes in the ODM.
   3. The MPIO scheduling algorithm may be changed to round\_robin if the reserve policy is changed to no\_reserve.
      1. chdev –l hdiskx –a algorithm=round\_robin
   4. Queue depth should be reviewed and increased in both the VIO servers and clients. Best practice is to make the queue depth the same in both the server and client partition. Queue\_depth limit for virtual scsi adapters is 512, of which 2 are used by the adapter.
   5. Review the hdisk attributes in the client partition and change the queue\_depth and hcheck\_interval.
   6. Path priority can be checked/changed for the hdisks in the VIO client to spread the workload.

To check the priority for each path, use the **lspath** command as follows:

# lspath -E -l hdisk1 -p vscsi0

priority 2 Priority True

# lspath -E -l hdisk1 -p vscsi1

priority 1 Priority True

See the IBM redbook on Virtual I/O – chapter 4.7.3 on configuring the VIO client: <http://www.redbooks.ibm.com/redpapers/pdfs/redp4194.pdf>

1. Virtual Fibre Channel (NPIV)
   1. HDLM or MPIO can be installed in the Client partition (not in the Virtual I/O servers). See the IBM redbook for instructions on implementing NPIV.
      1. <http://www.redbooks.ibm.com/redbooks/pdfs/sg247590.pdf>
   2. The reserve policy for HDLM must be changed to no\_reserve to use Live Partition Mobility.
   3. The HDLM NPIV option **must** be enabled.
      1. # /usr/D\*/bin/dlmodmset -o

Lun Reset : off

Online(E) IO Block : off

NPIV Option : off

KAPL10800-I The dlmodmset utility completed normally.

If you want to use HDLM in a client partition to which a virtual HBA is applied by using the virtual I/O server NPIV functionality, set the NPIV option to on. If the option is not set to on, HDLM might not be able to recognize a path that goes through the virtual HBA. Use the “dlmodmset –v on” before configuring the hdisks.

# Multipathing and Disk Considerations:

1. HDLM
   1. Recommended settings:
      1. Path Health Check = default value 30 minutes
      2. Auto Failback= On, interval between 5-15 minutes

*# dlnkmgr set –afb on –intvl 5*

* + 1. Intermittent Error Monitor= On, default value

# dlnkmgr set –iem on

* + 1. Load balancing= On, default value extended least I/O’s

#*# dlnkmgr set –lb on –lbtype exlio*

* + 1. Lun Reset On – Enables Adapter to identify failed path and to switch I/O**’s**

*#dlmodmset -r on (requires a reboot)*

* 1. Number of paths:
     1. 2 recommended
     2. 4 maximum – sufficient for availability and to reduce overhead of path selection.
  2. Change default hdisk attributes prior to discovering the disks
     1. Queue\_depth

*# dlmchpdattr –a queue\_depth=32*

* + 1. Reserve\_policy for VIOS or boot disks

*# dlmchpdattr –a reserve\_policy=no\_reserve*

* 1. Change NPIV option if required

*# dlmodmset –v on*

1. MPIO
   1. Settings
      1. Default settings: Single path reserve, failover only
      2. Can be modified to: No\_reserve, round\_robin

*# chdev –l hdiskx –a reserve\_policy=no\_reserve –a algorithm=round\_robin*

*These settings are required when using dual VIO servers or HACMP*

*Note: Care should be taken when using these settings as the reserves are disabled. Good SAN zoning practices are critical.*

* + 1. For AIX 6.1TL08 and AIX 7.1TL01 use the chdef command to permanently change attributes in the ODM predefined attributes (PdAt).

Usage:

chdef [-a Attribute=Value -c Class -s Subclass -t Type]

# odmget PdDv | grep -p htcvspmpio

PdDv:

type = "htcvspmpio"

class = "disk"

subclass = "fcp"

# chdef –a reserve\_policy=no\_reserve –c disk –s fcp –t htcvspmpio

Note: this could prevent Reservation Conflicts by setting default reserve policy to no\_reserve.

* + 1. New command “devrsrv” also available at the same AIX levels as chdef to display and

clear reserves.

Usage :

For Command Line Inputs:

devrsrv -c query | release | prin -s sa |

(prout -s sa -r rkey -k sa\_key -t prtype) -l devicename

devrsrv -f -l devicename

* + 1. New attribute, timeout\_policy with MPIO ODM update 5.4.0.4.
       - Effective in AIX 6.1TL06 and AIX 7.1
       - Recommended setting fail\_path
       - <http://www-01.ibm.com/support/docview.wws?uid=isg1IZ96396>

The timeout\_policy set to "fail\_path" will resolve many continuous performance degradation issues on MPIO devices connected via a path (or switch) going up and down.

* + 1. **Note: When changing the algorithm, ensure new disk definitions are changed to no\_reserve prior to discovering the disks. If the algorithm (global setting, part of the PCM definition) is round\_robin, the new disk definitions will have a default reservation setting of single\_path which will cause a conflict and prevent the new array’s disks from becoming available.**
  1. Number of paths:
     1. 2 recommended
     2. 4 maximum – sufficient for availability and to reduce overhead of path selection.