7135 RAIDiant Array: Install Issues for Twintailing

Special Notices

Information in this document is correct to the best of our knowledge at the time of this writing. Please send feedback by fax to "AIXServ Information" at (512) 823-4009.

Please use this information with care. IBM will not be responsible for damages of any kind resulting from its use. The use of this information is the sole responsibility of the customer and depends on the customer's ability to evaluate and integrate this information into the customer's operational environment.
H2.About This Document

This document contains setup tips for the IBM RAIDiant Array 7135 when shared by two or more systems. This document applies to AIX 3.2.5 and AIX 4.1. Parts of this document also apply to disk drives that accept SCSI commands and are shared by 2 or more systems.

Twin-Tailing the 7135 Between Multiple RISC System/6000s

Twin-tailing is hooking up a device to more than one system using a SCSI bus such that at least two SCSI adapters are on the same bus. It is possible to have more than two systems, more than two 7135s or more than two 7135 disk array controllers (dacs) on the same SCSI bus. This is limited by the SCSI standard limit of 8 devices on a SCSI bus, including the SCSI adapters. This is further limited in that a 7135 can be shared by no more than 4 systems.

If a host is twin-tailed to a 7135 with two dacs, it must be connected to both dacs. Connecting one dac to one machine and the other dac to the other machine will not work.

The SCSI-2 Fast/Wide Differential Adapter (Type 4-6) has the advantage of a maximum SCSI bus length of 25 meters, whereas the SCSI-2 DE (narrow) adapter (Type 4-2) is limited in bus length to 19 meters. Note that the 7135-210 requires use of the Enhanced SCSI-2 Fast/Wide Differential Adapter (Type 4-C) which has a maximum bus length of 25 meters.

The SCSI-2 Fast/Wide adapters actually have two SCSI busses: internal and external; consequently, it is configured as an "ascsi" device for the adapter and two "vscsi" devices, one for the internal bus and one for the external bus. You might also have a "tmscsi" device if you have the appropriate software installed. A tmscsi device is used for communication between systems using the SCSI bus as the physical medium (and is often used by HACMP).

The 7135 dacs are dual-ported differential SCSI devices, thus there are four SCSI ports per 7135. The bottom two ports connect to the dac in the back of the 7135 and the top two ports connect to the optional dac that goes in the front.

Device Names

It's good practice to setup twin-tailing such that the device names on both systems actually refer to the same device. This will help the system administrator from getting confused. Planning ahead is the key requirement to assure that the names are consistent on each machine. Remember that when AIX "walks the bus" and configures devices into the system, it starts with the lowest numbered adapter slots, then for SCSI adapters, starts configuring the device with SCSI ID 0, then 1 and so on. As AIX encounters a device, if it isn't already defined in the ODM database, a name is generated for the device using the next available number, e.g., if only one disk is in the system and used by rootvg (hdisk0), and a new disk is detected, it will be named hdisk1.

For hdisks, if both systems do not have the same number of internal disks, you can create a dummy or placeholder hdisk on the system with fewer disks to even them out. The easiest way to do this is to use the "smit makdsk" fastpath to add a disk, but before actually pressing enter to run the command, press F6 to get the command. Then run the command from the command line with an additional flag, "-d", to make the dummy device and leave it in the Defined state.

The next consideration is choice of adapter slots. We recommend that all adapters on the same SCSI bus be in the same slot on all systems. Doing this will assure that the dac(s), dar(s) and 7135 hdisk(s) are configured in the same order; thus having the same names on the twin-tailed systems. Finally, if you have more than one 7135 with two SCSI busses connected to each, we recommend that the adapters for each 7135 be in adjacent slots, then you will have dac0 and dac1 associated with dar0, dac2 and dac3 associated with dar1, etc. The exception would be if you have two Microchannel busses: then connecting adapters in slots 4 and 14 (to slots 4 and 14 in the twin-tailed system) could provide some availability and performance benefits; however, keeping track of which dac is associated with which dar will be more difficult.

Cabling

We highly recommend using Y cables when twin-tailing. The advantage of using Y cables is that you can disconnect a Y cable from a SCSI adapter and maintain termination on both ends of the SCSI bus. Thus you can perform hardware maintenance on the disconnected system without having to power down the other system(s). If you do not use Y cables, removing the cable from the adapter breaks the SCSI bus by leaving termination at only one end; thus you must have all devices on the bus powered off. Doing otherwise risks losing and/or corrupting existing data.

Both ends of the bus must be terminated, and using Y cables allows you to use terminators on the external cabling, rather than using terminators on the SCSI card. When using Y cables, you must remove the internal terminating resistors from the SCSI adapters. Refer to the "Common Diagnostics and Service Guide" SA23-2687 for instructions on removing all the internal terminating resistors.

Depending upon your situation, you might benefit from physically having the 7135 at the end of the SCSI bus rather than the middle (e.g., use of shorter cables) as pictured in the "7135 RAIDiant Array, Installation and Service Guide" SY33-0181-05. This is especially true when you are connecting the 7135 to SP2 nodes.

Twin-tailed SCSI buses should be quiesced prior to powering up/down or connecting/disconnecting either the RS/6000 or 7135 from the SCSI bus. These actions change the electrical characteristics of the SCSI bus and can interfere with transmissions on it. If data is being transferred, and two or more bits are changed such that the parity bit matches the data (thus the device drivers will not detect the error and handle it), you could have a byte of data corrupted. The chances of this are low; nevertheless, it is possible.

Make sure every device on the SCSI bus has a unique SCSI ID, including the adapters. By default, SCSI adapters have their SCSI ID set to 7, the recommended SCSI IDs are 5 and 6 for twin tailing. The reason for this is that if you reinstall from a backup or boot in maintenance mode, the SCSI ID will be set to 7, thus using 5 and 6 will prevent two adapters from having the same ID after a reinstall or during maintenance. For simplicity, we recommend all twin-tailed adapters on a system have the same SCSI ID, e.g, on system A set the IDs to 5 and on system B set the IDs to 6. To change the SCSI ID of an adapter, use the following procedure:

1. List all your SCSI devices and 7135 devices with

   `"lscfg | grep scsi" and "lsdev -C | grep Array"`

2. Varyoff any 7135 volume groups.
3. Run "rmdev -l <device>" for each device attached to the SCSI adapter, including hdisks, dars, dacs. For the Fast/Wide adapters, the devices also include the vscsi and tmscsi devices. This command will put the devices in the Defined state.
4. For the narrow adapter run "chdev -l <scsi#> -a id=5" where scsi# refers to the SCSI adapter of interest.
5. For the Fast/Wide adapter, run

   `chdev -l <ascsi#> -a external_id=6`

   where ascsi# is the adapter device of interest.
6. Run "cfgmgr" to put all the devices back in the Available state (or "ace -z dac0" for scarray.obj 3.2).

LVM Procecures

In addition to making all the hdisks Available on both systems, you must also make sure the physical volume IDs (PVIDs) are consistent on both systems. When you create a LUN/hdisk on one system, the second system will not be aware of it until you reboot or run cfgmgr. You should also be aware that after creating the LUN, the PVID will be set to all zeros. In this case, when you create a VG (or use "mkdev" for the hdisk), a PVID is generated, placed on the disk and in the ODM. On the other system; however, the ODM is not updated with the new PVID. To see if they are consistent, run the "lspv" command on both systems. If they are not consistent, run "rmdev -dl <hdisk#>" for the disk and then run "cfgmgr" to reconfigure the disk and update the ODM. You must be aware of the SCSI reserve when performing these operations (see the next section).

Without HACMP concurrent access, a volume group is active (varied on) on only one system at a time. You can always NFS mount it to the other system. While ways exist to have a VG active on two systems at once, insuring data integrity requires software that provides communication between the file locking mechanisms running on each system.

LVM does allow you to take a VG active on one system, deactivate it and make it active on a second twin-tailed system as follows:

On the first system:

1. Unmount the filesystems and deactivate (varyoff) the volume group (VG).
2. Export the VG - this is optional but recommended.

   On the second system:

3. Import the VG - this is optional but recommended. It is not optional if the VG was previously exported. Importing the VG assures that the information about the VG, including all it's logical volumes and filesystems, are accurate in the ODM and /etc/filesystems file.
4. Activate the VG and mount the filesystems.

One thing that may change about a VG when exporting and importing it is its quorum checking and automatic activation characteristics. These are always set to yes when they are imported. It is not recommended to automatically varyon twin-tailed VGs, especially when you boot both systems simultaneously. Thus you may want to run "chvg -a n <VGname>" to ensure the VG is not automatically varied on.

When not using HACMP, it is best not to have the same VG imported on both systems because if a) the VG is automatically activated and b) both systems are booted simultaneously, it is possible that both systems will varyon the VG - this could lead to data loss. We recommend only importing VGs that will be active on the system when not using HACMP.

If you do have the VG imported on both systems (e.g., if you are using HACMP) be aware that changes to any of the LVs (e.g. increasing the size of a filesystem or making a LV copy) must be propagated to both systems. Thus if you make a change on system A, you should export then reimport the VG on system B.

SCSI Reserve

A SCSI device can be reserved for a specific SCSI adapter. In this situation, the device will ignore requests from other adapters on the bus. Note that the SCSI device holds the reserve, not the SCSI adapter requesting it. AIX issues the SCSI reserve request when a) a volume group is activated and b) when reading the disk's physical volume ID (PVID) during boot. AIX tells the device to release the reserve when the VG is deactivated or after the PVID is read during boot.

If the computer/adapter that got the SCSI reserve loses power, the SCSI device will still hold the reserve. There are three ways for a 7135 to release a SCSI reserve: 1) The computer and adapter regains power and then a SCSI reserve release command is issued (e.g., varyoffvg), 2) Power cycle the 7135 and 3) using HACMP.

Booting twin-tailed RS/6000s simultaneously may create ghost disks. Booting a RS/6000 while the other one has twin-tailed VGs varied on will create ghost disks. Following is an example of a ghost disk:

```
# lsdev -Cc disk | grep  7135
  hdisk3 Available 00-03-00-00 7135 Disk Array Device
  hdisk4 Available 00-03-00-01 7135 Disk Array Device
  hdisk5 Defined   00-04-00-12 7135 Disk Array Device
  hdisk6 Available 00-03-00-03 7135 Disk Array Device
  hdisk7 Available 00-04-00-12 7135 Disk Array Device
```

Note that hdisk5 and hdisk7 have the same location code, but hdisk5 is in the Defined state. hdisk7 is the ghost disk and does not really exist. hdisk5 is in an active VG on the other twin-tailed RS/6000. The lspv command will not show VG membership for either hdisk5 or hdisk7, and it will not show a PVID for hdisk7.

Since hdisk5 is being used on the other system, there is nothing you can do with it until the other system varies off the VG. However, you can remove the ghost disk with "rmdev -dl hdisk7". To clean up the ghost disk after deactivating the hdisk5 VG on the other system, you would run "mkdev -l hdisk5".

If you must boot your systems simultaneously and you want your VG automatically varied on, you can use the following script, to delay the varyon command and give the other system a chance to configure the hdisk prior to the activation of the VG. The example script is named /usr/local/bin/delay_raid_varyon and follows:

```
#!/bin/ksh
# delay_raid_varyon - script to delay 7135 VG activation
# Use your own VG name instead of <raidvg>, and insure that
# the sleep time is sufficient.
  sleep 180
  varyonvg <raidvg>
  print "<raidvg> Activated" > /dev/console
```

Make the script executable by root, and then add the script to /etc/inittab with the following command:

```
mkitab 7135delay:2:once&colon/usr/local/bin/delay_raid_varyon
```

One method of dealing with ghost disks we discourage is setting the "Reservation Lock" parameter in the "Change/Show SCSI RAID Array" to no. While this will prevent ghost disk creation, it allows situations that can corrupt your data. This parameter is not available in the 3.2 (active-passive) version of the 7135 software.

Other Gotchas

Using dynamic load balancing with twin-tailing is not supported and will not work properly.

When twin-tailing, for scarray.obj 3.2, do not run any combination of diagnostics against a dac or the array configuration editor (ace) on more than one system at a time. Doing either of these can cause the systems to hang.

If you plan to use twin-tailing, or perhaps plan to implement high availability with HACMP, you can get assistance from AIX Support Family's Consult Line service.

Reader's Comments

Please fax this form to (512) 823-4009, attention "AIXServ Information". You may also e-mail comments to: elizabet@austin.ibm.com. These comments should include the same customer information requested below.

Use this form to tell us what you think about this document. If you have found errors in it, or if you want to express your opinion about it (such as organization, subject matter, appearance) or make suggestions for improvement, this is the form to use.

If you need technical assistance, contact your local branch office, point of sale, or 1-800-CALL-AIX (for information about support offerings). These services may be billable. Faxes on a variety of subjects may be ordered free of charge from 1-800-IBM-4FAX. Outside the U.S. call 415-855-4329 using a fax machine phone.

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you.

Note: If you have a problem report or item number, supplying that number may help us determine why a procedure did or did not work in your specific situation.

Problem Report or Item #: Branch Office or Customer #:

Be sure to print your name and fax number below if you would like a reply:

Name: Fax Number:

END OF DOCUMENT (7135.twintail.ha_, 4FAX# 2546)