Linux DPT Hardware RAID HOWTO

Table of Contents

Linux DPT Hardware RAID HOWTO

Ram Samudrala (me@ram.org)

v1.62, August 4, 2004

How to set up hardware RAID under Linux. This HOWTO is now limited to making small changes from version 1.6.

---

8. References

9. Acknowledgements

---

1. Introduction

This document describes how to set up SCSI hardware RAID, focusing mainly on host−based adapters from DPT, though the principles applied here are fairly general.

Use the information below at your own risk. I disclaim all responsibility for anything you may do after reading this HOWTO. The latest version of this HOWTO will always be available at http://www.ram.org/computing/linux/dpt_raid.html.

For the purposes of this HOWTO, I am assuming you have only a Linux system running. Also, note that I've only tried this out with the DPT Smartcache IV PM2144UW and PM3334UW controllers, with DPT (SmartRAID tower) and Wetex enclosures, and I have no experience with other setups. So things may be different for your setup.

---

2. Supported controllers

One well−supported host−based hardware RAID controller (i.e, a controller for which there exists a driver under Linux) is one that is made by DPT. However, there exist other host−based and SCSI−to−SCSI controllers which may work under Linux. These include the ones made by Syred, ICP−Vortex, and BusLogic. See the RAID solutions for Linux page for more info.

If, in the future, there is support for other controllers, I will do my best to incorporate that information into this HOWTO. Please send me any such information you think is appropriate for this HOWTO.

2.1 DPT controllers

This document is currently DPT−oriented. Essentially all the SmartRAID IV controllers are supported.

2.2 ICP vortex controllers

ICP vortex has a complete line of disk array controllers which support Linux. The ICP driver is in the Linux kernel since version 2.0.31. All major Linux Distributors S.u.S.e., LST Power Linux, Caldera and Red Hat support the ICP controllers as boot/installation controllers. The RAID system can easily be configured with their ROMSETUP (you do not have to boot MS−DOS for configuration!).

With the monitoring utility GDTMON it is possible to manage the complete ICP RAID system during operation (check transfer rates, set parameters for the controller and hard disks, exchange defective hard disks, etc.). Currently available are: 1 and 2 channel wide and ultra SCSI controller for RAID 0 and RAID 1 1, 2, 3 and 5 chn. wide and ultra SCSI controller for RAID 0, 1, 4, 5 and 10 1 and 2 channel wide and ultra2 LVDS SCSI controller for RAID 0 and RAID 1 1, 2, 3 and 5 chn. wide and ultra2 LVDS SCSI controller for RAID 0, 1, 4, 5 and 10 1 and 2 port Fibre Channel controllers for RAID 0, 1, 4, 5 and 10 Pretty soon there will be also 64−bit controllers available.

ICP is transitioning the entry−level RS series from Ultra2 SCSI to Ultra160 SCSI. The drivers, firmware, features, capabilities etc remain the same. They are still 32 Bit cards with the i960RS processor working at 100MHz. The only difference is they will work at Ultra160 (data transfer rate of 160MB/sec) rather than Ultra2 (data transfer of 80MB/sec).

Effective immediately, the GDT7523RN units will become GDT8523RZ and the GDT7623RN units will become GDT8623RZ. The transition from 33MHz on the PCI bus to 66MHz represents a huge potential performance increase. The new cards will have the new Intel 80303 "Zion" processor, allowing bus master transfer rates of up to 528MB/sec, and will take up to 256MB of ECC RAM on PC133 SDRAM Dimms.

3. What hardware should be used?

3.1 Controller type

Given all these options, if you're looking for a RAID solution, you need to think carefully about what you want. Depending on what you want to do, and which RAID level you wish to use, some cards may be better than others. SCSI−to−SCSI adapters may not be as good as host−based adapters, for example. Michael Neuffer (neuffer@uni−mainz.de), the author of the EATA−DMA driver, has a nice discussion about this on his Linux High Performance SCSI and RAID page.

3.2 Enclosure type

The enclosure type affects the hot swap−ability of the drive, the warning systems (i.e., whether there will be indication of failure, and whether you will know which drive has failed), and what kind of treatment your drive receives (for example, redundant cooling and power supplies). We used the DPT supplied enclosures which work extremely well, but they are expensive.

4. Installation

4.1 Installing and configuring the hardware

Refer to the instruction manual to install the card and the drives. For DPT, since a storage manager for Linux doesn't exist yet, you need to create a MS−DOS−formatted disk with the system on it (usually created using the command "format /s" at the MS−DOS prompt). You will also be using the DPT storage manager for MS−DOS (available from the Adaptec website), which you should probably make a copy of for safety.

Once the hardware is in place, boot using the DOS system disk. Replace the DOS disk with the storage manager. And invoke the storage manager using the command:

```
a:\ dptmgr
```

Wait a minute or so, and you'll get a nice menu of options. Configure the set of disks as a hardware RAID (single logical array). Choose "other" as the operating system.

The MS−DOS storage manager is a lot easier to use with a mouse, and so you might want to have a mouse driver on the initial system disk you create.

Technically, it should be possible to run the SCO storage manager under Linux, but it may be more trouble than its worth. It's probably more easier to run the MS−DOS storage manager under Linux.

4.2 Configuring the kernel

You will need to configure the kernel with SCSI support and the appropriate low level driver. See the Kernel HOWTO for information on how to compile the kernel. Once you choose "yes" for SCSI support, in the low level drivers section, select the driver of your choice (EATA DMA or EATA ISA/EISA/PCI for most EATA DMA compliant (DPT) cards, EATA PIO for the very old PM2001 and PM2012A from DPT). Most drivers, including the EATA DMA and EATA ISA/EISA/PCI drivers, should be available in recent kernel versions.

Once you have the kernel compiled, reboot, and if you've set up everything correctly, you should see the driver recognising the RAID as a single SCSI disk. If you use RAID−5, you will see the size of this disk to be 2/3 of the actual disk space available.

4.3 Bootup messages

The messages you see upon bootup if you're using the EATA DMA driver should look something like this:

```
EATA (Extended Attachment) driver version: 2.59b
developed in co−operation with DPT
(c) 1993−96 Michael Neuffer, mike@i−Connect.Net
Registered HBAs:
HBA no. Boardtype    Revis  EATA Bus  BaseIO IRQ DMA Ch ID Pr QS  S/G IS
scsi0 : PM2144UW     v07L.Y 2.0c PCI  0xef90  11 BMST 1  7  N  64 252 Y
scsi0 : EATA (Extended Attachment) HBA driver
scsi : 1 host.
  Vendor: DPT       Model: RAID−5            Rev: 07LY
  Type:   Direct−Access                      ANSI SCSI revision: 02
Detected scsi disk sda at scsi0, channel 0, id 8, lun 0
scsi0: queue depth for target 8 on channel 0 set to 64
scsi : detected 1 SCSI disk total.
SCSI device sda: hdwr sector= 512 bytes. Sectors= 35591040 [17378 MB] [17.4 GB]
```

(The above display is for a setup with a single DPT SCSI controller, configured as RAID−5, with three disks of 9 GB each.)

The messages you see upon bootup if you're using the EATA ISA/EISA/PCI driver should look something like this:

```
aic7xxx: <Adaptec AHA−294X SCSI host adapter> at PCI 15
aic7xxx: BIOS enabled, IO Port 0x7000, IO Mem 0x3100000, IRQ 15, Revision B
aic7xxx: Single Channel, SCSI ID 7, 16/16 SCBs, QFull 16, QMask 0x1f
EATA0: address 0x7010 in use, skipping probe.
EATA0: 2.0C, PCI 0x7410, IRQ 11, BMST, SG 252, MB 64, tc:y, lc:y, mq:62.
EATA0: wide SCSI support enabled, max_id 16, max_lun 8.
EATA0: SCSI channel 0 enabled, host target ID 6.
EATA/DMA 2.0x: Copyright (C) 1994−1997 Dario Ballabio.
scsi0 : Adaptec AHA274x/284x/294x (EISA/VLB/PCI−Fast SCSI) 4.1.1/3.2.1
scsi1 : EATA/DMA 2.0x rev. 3.11.00
scsi : 2 hosts.
scsi0: Scanning channel A for devices.
  Vendor: IBM OEM   Model: DFHSS2F           Rev: 1818
  Type:   Direct−Access                      ANSI SCSI revision: 02
Detected scsi disk sda at scsi0, channel 0, id 0, lun 0
```

```
  Vendor: SEAGATE   Model: ST41650       TX  Rev: DG01
  Type:   Direct-Access                      ANSI SCSI revision: 02
Detected scsi disk sdb at scsi1, channel 0, id 0, lun 0
  Vendor: TEAC      Model: FC-1     GF   00  Rev: RV L
  Type:   Direct-Access                      ANSI SCSI revision: 01 CCS
Detected scsi removable disk sdc at scsi1, channel 0, id 3, lun 0
  Vendor: SONY      Model: CD-ROM CDU-541    Rev: 2.6a
  Type:   CD-ROM                             ANSI SCSI revision: 02
Detected scsi CD-ROM sr0 at scsi1, channel 0, id 5, lun 0
EATA0: scsi1, channel 0, id 0, lun 0, cmds/lun 21, sorted, tagged.
EATA0: scsi1, channel 0, id 3, lun 0, cmds/lun 21, sorted.
EATA0: scsi1, channel 0, id 5, lun 0, cmds/lun 21, sorted.
scsi : detected 1 SCSI cdrom 3 SCSI disks total.
SCSI device sda: hdwr sector= 512 bytes. Sectors= 4404489 [2150 MB] [2.2 GB]
SCSI device sdb: hdwr sector= 512 bytes. Sectors= 2779518 [1357 MB] [1.4 GB]
SCSI device sdc: hdwr sector= 256 bytes. Sectors= 4160 [1 MB] [0.0 GB]
```

(The above display is for a setup with two SCSI controllers, DPT PM3224W and and Adaptec AHA2940.)

5. Usage

5.1 fdisk, mke2fs, mount, etc.

You can now start treating the RAID as a regular disk. The first thing you'll need to do is partition the disk (using fdisk). You'll then need to set up an ext2 filesystem. This can be done by running the command:

```
% mkfs -t ext2 /dev/sdxN
```

where /dev/sdxN is the name of the SCSI partition. Once you do this, you'll be able to mount the partitions and use them as you would any other disk (including adding entries in /etc/fstab).

5.2 Hot swapping

We first tried to test hot swapping by removing a drive and putting it back in the DPT−supplied enclosure/tower (which you buy for an additional cost). Before we could carry this out to completion, one of the disks failed (as I write this, the beeping is driving me crazy). Even though one of the disks failed, all the data on the RAID drive was accessible.

Instead of replacing the drive, we just went through the motions of hot swapping and put the same drive back in. The drive rebuilt itself and everything turned out okay. During the time the disk had filed, and during the rebuilding process, all the data was accessible. Though it should be noted that if another disk had failed, we'd have been in serious trouble.

5.3 Performance

Here's the output of the Bonnie program, on a 2144 UW with 9x3=17 GB RAID 5 setup, using the EATA DMA driver. The RAID is on a dual processor Pentium Pro machine running Linux 2.0.33. For comparison, the Bonnie results for the IDE drive on that machine are also given.

```
           -------Sequential Output-------- ---Sequential Input--  --Random--
           -Per Char- --Block--- -Rewrite-- -Per Char- --Block---  --Seeks---
        MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU   /sec  %CPU
```

```
RAID   100  9210 96.8  1613  5.9   717  5.8  3797 36.1 90931 96.8 4648.2 159.2
IDE    100  3277 32.0  6325 23.5  2627 18.3  4818 44.8 59697 88.0  575.9  16.3
```

Some people have disputed the above timings (and rightly so−−−I've been unable to try it out on our machines since they're completely loaded) because the size of the file used may have led to it being cached (resulting in an unusually good performance report). Here are some timings with a 3344 UW controller:

```
        −−−−−−−Sequential Output−−−−−−−− −−−Sequential Input−− −−Random−−
        −Per Char− −−Block−−− −Rewrite−− −Per Char− −−Block−−− −−Seeks−−−
     MB K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU  /sec %CPU
   1000  1714 17.2  1689  6.0  1200  5.7  5263 40.2  7023 12.1  51.3 2.2
```

And here are some timings on a SCSI−to−SCSI RAID system:

```
     −−−−−−−Sequential Output−−−−−−−− −−−Sequential Input−− −−Random−−
     −Per Char− −−Block−−− −Rewrite−− −Per Char− −−Block−−− −−Seeks−−−
MB   K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU K/sec %CPU /sec %CPU
64   7465 100.0 70287 98.7 37012 97.7 8074 99.2 *****100.3 ***** 196.6
128  7289 99.3  67595 98.5 35294 98.6 7792 97.6 *****100.3 ***** 195.8
256  7222 98.8  44844 69.6 16096 51.8 5787 72.7 ***** 99.8 ***** 85.2
512  7138 98.4  13871 23.2 7888  29.3 7183 89.3 16488 27.2 1585. 11.5
1024 6908 95.8  12270 21.5 7161  25.4 7373 90.4 16527 28.2 123.8 1.8
2047 6081 84.1  12664 22.6 7191  25.6 7289 89.5 16573 28.5 75.0  1.2

***** results exceed column width (> 100 MB/sec, > 10000 seeks/sec)

host:   Dual PII 400 MHz, 2 x U2W, 512 MB RAM, no internal disks
RAID:   IFT 3102 UA 128 MB Cache, RAID−5, 6 x 9 GB
OS:     SuSE Linux 6.0 with Kernel 2.2.3
```

6. Features in the EATA DMA driver

This section describes some of the commands available under Linux to check on the RAID configuration. Again, while references to the eata_dma driver is made, this can be used to check up on any driver.

To see the configuration for your driver, type:

```
% cat /proc/scsi/eata_dma/N
```

where N is the host id for the controller. You should see something like this:

```
EATA (Extended Attachment) driver version: 2.59b
queued commands:         353969
processed interrupts:    353969

scsi0 : HBA PM2144UW
Firmware revision: v07L.Y
Hardware Configuration:
IRQ: 11, level triggered
DMA: BUSMASTER
CPU: MC68020 20MHz
Base IO : 0xef90
Host Bus: PCI
SCSI Bus: WIDE Speed: 10MB/sec.
SCSI channel expansion Module: not present
SmartRAID hardware: present.
    Type: integrated
```

```
    Max array groups:              7
    Max drives per RAID 0 array:   7
    Max drives per RAID 3/5 array: 7
Cache Module: present.

    Type: 0
    Bank0: 16MB without ECC
    Bank1: 0MB without ECC
    Bank2: 0MB without ECC
    Bank3: 0MB without ECC
Timer Mod.: present
NVRAM     : present
SmartROM  : enabled
Alarm     : on
Host<->Disk command statistics:
         Reads:      Writes:
    1k:           0            0
    2k:           0            0
    4k:           0            0
    8k:           0            0
   16k:           0            0
   32k:           0            0
   64k:           0            0
  128k:           0            0
  256k:           0            0
  512k:           0            0
 1024k:           0            0
>1024k:           0            0
Sum   :           0            0
```

To get advanced command statistics, type:

```
% echo "eata_dma latency" > /proc/scsi/eata_dma/N
```

Then you can do a:

```
% cat /proc/scsi/eata_dma/N
```

to get more detailed statistics.

To turn off advanced command statistics, type:

```
% echo "eata_dma nolatency" > /proc/scsi/eata_dma/N
```

7. Troubleshooting

7.1 Upon bootup, no SCSI hosts are detected

This could be due to several reasons, but it's probably because the appropriate driver is not configured in the kernel. Check and make sure the appropriate driver (EATA−DMA or EATA ISA/EISA/PCI for most DPT cards) is configured.

7.2 RAID configuration shows up as N different disks

The RAID has not been configured properly. If you're using a DPT storage manager, you need to configure the RAID disks as a single logical array. Michael Neuffer (neuffer@uni−mainz.de) writes: "When you configure the controller with the SM start it with the parameter /FW0 and/or select Solaris as OS. This will cause the array setup to be managed internally by the controller."

7.3 Machine/controller is shut down in the middle of a format

As stated in the DPT manual, this is clearly a no−no and might require the disks to be returned to the manufacturer, since the DPT Storage Manager might not be able format it. However, you might be able to perform a low level format on it, using a program supplied by DPT, called clfmt in their utilities page. Read the instructions after unzipping the clfmt.zip file on how to use it (and use it wisely). Once you do the low level format, you might be able to treat the disks like new. Use this program carefully!

7.4 SCSI_ABORT_BUSY errors produced during initial filesystem format

When you do a `mke2fs` on the SCSI drive, you may see errors of the form:

```
scsi: aborting command due to timeout : pid xxx, scsi0, channel 0, id
2, lun 0
write (10) xx xx xx xx xx xx xx xx xx
eata_abort called pid xxx target: 2 lun: 0 reason: 3
Returning: SCSI_ABORT_BUSY
```

and this might end up causing the machine to freeze. I (and many others) have been able to fix this problem by simply reading one or two hundred MB from the RAID array with `dd` like this:

```
% dd if=/dev/sdX of=/dev/null bs=1024k count=128
```

During a format, a fast rush of requests for chunks of memory that is directly accessible is made, and sometimes the memory manager cannot deliver it on time anymore. The `dd` is a workaround that will simply create the requests sequentially instead of one huge heap at once like the format tends to create it.

7.5 If all fails...

Read the SCSI−HOWTO again. Check the cabling and the termination. Try a different machine if you have access to one. The most common cause of problems with SCSI devices and drivers is because of faulty or misconfigured hardware. Finally, you can post to the various newsgroups or e−mail me, and I'll do my best to get back to you.

---

8. References

The following documents may prove useful to you as you set up RAID:

- DPT Technology Library

- EATA−DMA homepage
- Linux Disk HOWTO
- Linux Kernel HOWTO
- Linux SCSI HOWTO
- Multi Disk System Tuning HOWTO
- RAID Solutions for Linux

---

9. Acknowledgements

The following people have been helpful in getting this HOWTO done:

- Andreas Koepf (A_Koepf@icp−vortex.com)
- Boris Fain (fain@zen.stanford.edu)
- Dario Ballabio (Dario_Ballabio@milano.europe.dg.com)
- Heiko Rommel (Heiko.Rommel@Uni−Bielefeld.DE)
- Jos Vos (jos@xos.nl)
- Michael Neuffer (neuffer@uni−mainz.de)
- Ralph Wallace (rwallace@rwallace.interaccess.com)
- Russell Brown (russell@lutton.lls.com)
- Syunsuke Ogata (Syunsuke_Ogata@appear.ne.jp)
- Tom Brown (tbrown@baremetal.com)
