

# Handleiding raid-1-server

Lieven Baes aka yanu

9 januari 2005

## Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>1</b>
<b>2</b>	<b>Installatie van raid.</b>	<b>1</b>
2.1	De linux-kernel. . . . .	1
<b>3</b>	<b>Installatie van het systeem.</b>	<b>2</b>
<b>4</b>	<b>Installatie van het raid-systeem.</b>	<b>3</b>
4.1	Partitioneren van de sata-schijven. . . . .	4
4.2	Aanmaken van de raid-arrays. . . . .	5
4.3	Kopiëren van een systeem. . . . .	6
4.4	Instellen van het systeem. . . . .	6
<b>5</b>	<b>Troubleshooting.</b>	<b>8</b>
5.1	Enkele commando's van raid. . . . .	8
5.2	Swap in raid-0. . . . .	10
5.3	Een nieuwe harde schijf erbij. . . . .	10
5.4	Het superblock van een array stemt niet overeen. . . . .	10
<b>6</b>	<b>Aanmaken van raid met 1 harde schijf en later de 2de toevoegen.</b>	<b>11</b>
<b>7</b>	<b>Een mdadm config file aanmaken.</b>	<b>12</b>
<b>8</b>	<b>Ondervonden problemen.</b>	<b>12</b>

## 1 Inleiding

De bedoeling is een machine op te zetten waarbij de gegevens weggeschreven worden op twee harde schijven, een redundant systeem.

Bij uitval van 1 harde schijf, doet de tweede gewoon verder.

De slechte harde schijf wordt eruitgenomen, een nieuwe erin, liefst van hetzelfde type, partities aanmaken, type zetten, en terug aankoppelen (ik geloof dat er geen filesystem moet aangemaakt worden, maar ik ben dat niet zeker).

Voor dit systeem werd gebruik gemaakt van twee identieke sata-schijven van 80Gig. Het besturingssysteem wordt Debian-testing (07-01-2005).

## 2 Installatie van raid.

### 2.1 De linux-kernel.

Het simpelste is zelf een kernel bakken, met daarin 0 en raid-1 support ingebakken. Dit heeft een nadeel dat je dan alle andere hardware zelf moet inbakken of als modules gebruiken. De grootte van de kernel is dan ook beperkt.

De raid-kernelsoftware kan bij het opstarten op twee manieren geladen worden, zodanig dat de raid-arrays (/dev/mdx) worden herkend.

- door de raid-soft in te bakken in de kernel.
- via initrd, dit is een initiële root-directory die aangemaakt wordt bij de opstart.

Met de default kernel van debian (kernel-image-2.6.8-1-686) werkte dit goed. In de testfase maakte ik gebruik van 1 idebus, de primaire. Er was geen secundaire aanwezig op die testmachine.

Om de performantie zo hoog mogelijk te houden is het best om de twee schijven op verschillende bussen te zetten. Eén ide-bus met twee harde schijven wil je niet in raid zetten, veel te traag.

De uiteindelijke machine had er wel twee waardoor 1 schijf op een ander ide werd gezet en dan werkte de initiële ramdisk niet meer.

Er moest dus een nieuwe initrd aangemaakt worden (mkinitrd -o /boot/initrd.img-2.6.8-1-686 2.6.8-1-686). Dit werkte niet op Debian. Na veel zoekwerk stootte ik op dit in de manpage van mkinitrd :

```
If both mdadm(8) and raidtools2 are installed, the former is preferred.
At the moment, mkinitrd uses the -D option of mdadm(8) to discover the
constituent devices. This means that only devices that are part of the
array at the time that mkinitrd is run will be used later on.
This problem does not exist when raidtools2 is used.
```

Weg default kernel dus. Zelf kernel bakken . . .

Met kernel 2.8.6 werkte de raid op zich wel goed, maar bij het syncen liep het systeem met sata-schijven geleidelijk aan vast. Dit was niet het geval met ide-schijven. Ik steek de schuld op ofwel de kernel2.6.8 ofwel de sata-driver (PIIX (ata\_piix) in dit geval ofwel de combinatie van de twee. Nergens op het net werd dit probleem vermeld.

Dan maar de kernel 2.4.28! Daarin bolt alles naar wens.

#### Opmerking:

Opdat *poweroff* de machine zou doen stilvallen, moest in de bios, *acpi apci enabled* zijn en *apm off*.

Ook moet *acpi* in de kernel meegecompileerd worden (alle andere modules heb ik gelaten zoals ze waren). *apm* heb ik als module gecompileerd en uitgeschakeld in */etc/lilo.conf*.

```
image=/boot/vmlinuz-2.6.8-acpi
label=Lx-2.6.8-acpi
#append="nolapic noapic"
append="noapm"
```

Daarnaast moet je ook nog het pakket *acpid*, de acpi-daemon installeren.

### 3 Installatie van het systeem.

Ik koos ervoor om 3 partities te maken over de twee schijven:

1. de swap in raid-0. Het is beter de swap in raid-0 te plaatsen.  
Een nadeel hiervan is, dat bij een diskuitval er bij het heropstarten geen swap-ruimte meer zal beschikbaar zijn (Uitleg zie verder in Troubleshooting).
2. de root-partitie in raid-1
3. de home-partitie in raid-1

Dit wordt d.m.v Knoppix gedaan. De uitleg vind je op <http://juerd.nl/site.plp/debianraid>.

Hierbij gaat men de twee harde schijven partitioneren, de raid-devices aanmaken, de devices formatteren (ik gebruik reiserfs), chrooten en Debian via debootstrap installeren, nieuwe kernel compileren, *lilo* configureren en de netwerkomgeving opzetten.

Alle zaken kan je naar eigen voorkeur bijstellen:

- woody (stabel) → sarge (testing)
- */target ftp://ftp.belnet/packages/debian*
- */usr/share/zoneinfo/Europe/Amsterdam* → */usr/share/zoneinfo/Europe/Brussels*

- */etc/apt/sources.list*
- kernel → kernel 2.4.28
- apt-get update && apt-get upgrade
- ....

Bij Debian kan je een aantal zaken juist zetten, zoals tijd, keyboard, ...

- dpkg-reconfigure console-data
- timezone met *tzconfig*
- tijdssynchronisatie instellen met *ntpdate* en een script in */etc/cron.daily*.

```
#!/bin/sh
# dit zet je pctijd gelijk met de ntp-server
# deze wordt dan ook de in hardwareclock geschreven
# eigenlijk "ntp.charon.telenet-ops.be"
# een wereldwijde server: pool.ntp.org
# zie http://www.die.net/doc/linux/HOWTO/Belgian-HOWTO/isp.html
# voor meer info

exec ntpdate ntp.charon.telenet-ops.be && exec hwclock --systemd
```

## 4 Installatie van het raid-systeem.

Dit werd eigenlijk al gedaan via de installatie van Debian, maar ik zet het hier nog eens stap voor stap neer, omdat je ook een bestaand systeem kunt kopiëren op een zelfgebouwd raid-systeem.

Ook hier maak ik gebruik van Knoppix. Knoppix start ik "bijnaãltijd op in console. In het boot-scherm geef je minimum dit mee:

```
knoppix26 2 lang=be
```

of

```
knoppix26 2 lang=be noswap
```

### Opmerking ivm Knoppix:

Knoppix is een cd-distro en zal al wat hij vindt gaan gebruiken, zoals bestaande swappartities. Dit kan je zien aan de laatste lijnen bij het opstarten.

Voordat je aan de raid begint, leg je ze best af.

```
Using swap partition /dev/sda1
Using swap partition /dev/sdb1

swapoff /dev/sda1
swapoff /dev/sdb1
```

Hetzelfde bereik je ook met de optie *noswap* bij het booten.

#### 4.1 Partitioneren van de sata-schijven.

Controle van de namen van de schijven, je verkrijgt iets in deze aard:

```
fdisk -l

Disk /dev/sda: 80.0 GB, 80026361856 bytes
255 heads, 63 sectors/track, 9729 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes

Disk /dev/sdb: 80.0 GB, 80026361856 bytes
255 heads, 63 sectors/track, 9729 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
```

Voor dit systeem heb ik 3 partities aangemaakt, ze worden later alledrie in raid gezet:

- *sda1* en *sdb1* als swap (1014MB)
- *sda2* en *sdb2* als root (2048MB)
- *sda3* en *sdb3* als home (data) (de rest van de schijf)

Het makkelijkste als je twee identieke schijven hebt, is de eerste schijf partitioneren en dan de mbr (master boot record) kopiëren naar de andere.

Waar moet je op letten:

- Zet de *root*-partities (*/dev/sda2* en */dev/sdb2*) bootable.  
Ik denk niet dat dit nodig is, nuja ...
- Dit is van groter belang!  
Zet **het id-type** van alle partities op *Linux raid autodetect* of *fd*.

De ene partitie-gegevens kopiëren naar de andere:

```
sfdisk -d /dev/sda | sfdisk /dev/sdb
```

Nu **moet** je rebooten opdat de kernel de juiste partitiegegevens zou inlezen en gebruiken.

## 4.2 Aanmaken van de raid-arrays.

Terug in Knoppix, de swap-partities eraf gooien of *noswap*.

De raid-arrays mag je beginnen tellen vanaf 0. Voor het gemak begin ik vanaf 1, dan komen de cijfers overeen met deze van de gebruikte partities.

**Let op het raid-level 0 voor /dev/md1.**

```
# mdadm --create /dev/md1 -n 2 -l 0 /dev/sda1 /dev/sdb1
# mdadm --create /dev/md2 -n 2 -l 1 /dev/sda2 /dev/sdb2
# mdadm --create /dev/md3 -n 2 -l 1 /dev/sda3 /dev/sdb3
```

*-n 2* betekent dat er twee disken gebruikt worden in die raid-array.

*-l 0* betekent dat raid-0 wordt gebruikt.

*-l 1* betekent dat raid-1 wordt gebruikt.

Wachten totdat ze gesynct zijn (*watch cat /proc/mdstats*).

Indien je ergens zou gemist zijn, dan kan je altijd de raid-array stilleggen en hermaken.

Best wel direct daarna de superblocs op 0 zetten (zie iets verder).

```
# mdadm -S /dev/md1
```

### Note:

Je kan ook een raid-systeem opzetten met 1 harde schijf. Later als alles klaar is, steek je de tweede erbij en koppel je hem.

Uitleg hierover vind je verder bij *Troubelshooting*.

De raid-arrays formatteren met het gewenste filesystem:

```
# mkswap /dev/md1
# mkreiserfs /dev/md2
# mkreiserfs /dev/md3
```

### Opmerking:

Normaal als je raid-arrays aanmaakt, (moet) je ze formatteren.

Ik geloof dat het eigenlijk niet echt verschil maakt of ze vooraf geformatteerd waren op de fysische devices (*/dev/sdx*). Het 'moet' dus **niet**.

Wat er wel van belang is, zijn de *superblocks* die de raid-arrays meekrijgen bij het aanmaken van de raid-array.

De superblocs moeten voor beiden overeenstemmende partities gelijk zijn, zodat de kernel ze kan ontdekken en samenvoegen tot 1 raid-array (bv */dev/md1*).

Zijn de *superblocks* niet gelijk, dan zet je ze op 0 en maak je ze opnieuw aan:

```
# mdadm --zero-superblock /dev/sda1
# mdadm --zero-superblock /dev/sda2
# mdadm --zero-superblock /dev/sda3
# mdadm --zero-superblock /dev/sdb1
# mdadm --zero-superblock /dev/sdb2
# mdadm --zero-superblock /dev/sdb3
```

En de raid-arrays terug aanmaken (-A).

Op deze manier verschijnen overeenstemmende *superblock*-waarden, ...

### 4.3 Kopiëren van een systeem.

Mount zowel de source- (waarop je systeem staat) als de target-directorie in Knoppix.

Bij mij staat de source als *root.tar.gz*-file op een cdrom. Via het *scp* verplaats ik die naar de rootmap (*/dev/md2*).

Het target is */dev/md2*

```
# mkdir -p /md2 /cdrom
# mount /dev/md2 /md2
# mount /dev/cdrom /cdrom
# cp /cdrom/root.tar.gz /md2/
# cd /md2
# tar zxvf root.tar.gz ./

# cd /md2/
# tar zxvf root.tar.gz ./
met mc op de juiste plaats zetten
```

De inhoud van de *home* naar */dev/md3* verplaatsen. */dev/md3* mount je op */md3* nadat je deze map hebt aangemaakt.

De *home* en *cdrom* unmounten (*md2*, */md3* en */cdrom*).

### 4.4 Instellen van het systeem.

1. Om het systeem te kunnen booten, moet je */etc/fstab* aanpassen zodat het bootproces de juiste arrays gaat mounten. */etc/fstab* ziet er in dit geval zo uit:

```
# This is /etc/fstab

/dev/md1 none swap sw 0 0
/dev/md2 / reiserfs defaults 0 1
/dev/md3 /home reiserfs defaults 0 2
proc /proc proc
```

**Opmerking:**

De laatste 1 betekent dat het filesystem voordat het gemount wordt, zal als eerste gecheckt worden op fouten. In dit geval zal *reiserfsck /dev/md2* uitgevoerd worden.

Staat er een 2, dan zullen alle andere als tweede gecheckt worden.

Bij reiserfs mag je dit zo doen.

2. Indien er device-namen veranderd zijn, moet */etc/lilo.conf* ook aangepast worden. */etc/fstab* ziet er in dit geval zo uit:

```
# This is /etc/lilo.conf

# boot from first raid-blockdevice
boot=/dev/md2
# root is on first raid-blockdevice
root=/dev/md2
# This writes the boot signatures to either disk
raid-extra-boot=/dev/sda,/dev/sdb

#compact
lba32
read-only
menu-title= " Sambaserver"
#map=/boot/map # enkel voor oudere machines, voor 1998
prompt
timeout=60 # 3sec wachten
#delay=50

# Enkel deze werkt bij het syncen van de raid
default=Lx-2.4.28-raid
#default=Lx-2.6.10-acpi
#default=Lx-2.6.8-acpi

# deze kernel werkt bij het syncen
image=/boot/vmlinuz-2.4.28-raid
label=Lx-2.4.28-raid
append="noapm"

# deze twee zijn reservekernels, ze werken niet bij sync
image=/boot/vmlinuz-2.6.10-acpi
label=Lx-2.6.10-acpi
#append="nolapic"
append="noapm"
image=/boot/vmlinuz-2.6.8-acpi
label=Lx-2.6.8-acpi
#append="nolapic"
append="noapm"

# dit test het geheugen
image=/boot/memtest
label=memtest86
```

Vergeet dan *lilo* niet te draaien!

Hiervoor moet je het root-filesysteem wel **chrooten** (= een linux-systeem onder linux draaien).

Om te chrooten, moet je in Knoppix tijdens het mounten de optie *-i dev* meegeven.

```
# mount -o dev /dev/md2 /md2
# chroot -o dev /md2
sh-2.05# lilo -t      (testen of alles juist is in /etc/lilo)

sh-2.05# lilo

sb-2.05# exit
```

Nu kan je het systeem rebooten en hopen dat de kernel de raid-arrays vindt en laadt.

## 5 Troubleshooting.

### 5.1 Enkele commando's van raid.

Het zichtbaar maken van de status van een raid-systeem:

```
# cat /proc/mdstat

Personalities : [raid1]
md1 : active raid1 sdb1[1] sda1[0]
      8225152 blocks [2/2] [UU]

md2 : active raid1 sdb2[1] sda2[0]
      8225216 blocks [2/2] [UU]

md3 : active raid1 sdb3[1] sda3[0]
      61697536 blocks [2/2] [UU]

unused devices: <none>
```

Voortdurend de status volgen:

```
# watch cat /proc/mdstat
Personalities : [linear] [raid0] [raid1] [raid5] [multipath]
read_ahead 1024 sectors
md1 : active raid0 sdb1[1] sda1[0]
      995712 blocks 64k chunks

md2 : active raid1 sdb2[1] sda2[0]
      2000000 blocks [2/2] [UU]

md3 : active raid1 sdb3[2] sda3[0]
```

```
75649984 blocks [2/1] [U_]
[==>.....] recovery = 13.1% (9985448/75649984) finish=21.5min speed
unused devices: <none>
```

Informatie vragen over een raid-device:

```
# mdadm -D /dev/md3
```

Een partitie als fout zetten (*-fail* of *-f*):

```
# mdadm /dev/md3 -f /dev/sdb3
```

Een partitie verwijderen (*-remove* of *-r*):

```
# mdadm /dev/md3 -r /dev/sdb3
```

Een partitie aankoppelen (*-add* of *-a*):

```
# mdadm /dev/md3 -a /dev/sdb3
```

Alles samendoen:

```
# mdadm /dev/md3 -f /dev/sdb3 -r /dev/sdb3 -a /dev/sdb3
```

Let op het syncen.

Wanneer je boot met Knoppix, om iets te herstellen bevoorbeeld, moet je zelf de raid opnieuw aanmaken (*assemble*: *-assemble* of *-A*).

```
# swapoff /dev/sda1
# swapoff /dev/sdb1
# mdadm -A /dev/md1 /dev/sda1 /dev/sdb1
# mdadm -A /dev/md2 /dev/sda2 /dev/sdb2
# mdadm -A /dev/md3 /dev/sda3 /dev/sdb3
```

Let op hoe ze gesynct worden.

Daarna kan je ze mounten en bewerken.

## 5.2 Swap in raid-0.

Door de swap in raid-0 te zetten, verkrijg je dubbele hoeveelheid swap, dus kan je evengoed je partitie wat verkleinen.

Wanneer er nu een disk uitvalt, moet de server verder werken zonder swap, aangezien de array een stuk mis. De kernel kan ze dus niet gebruiken.

Met het commando *free* zie je of de swap gebruikt wordt.

Je kunt de swap-ruimte van de goeie disk aanwenden om het systeem verder te laten werken op 1 schijf. Dit moet je wel handmatig doen.

**Opgelet! dat je geen fouten maakt met de nummers.**

```
# swapon /dev/sda1
```

Wanneer je beschikt over een nieuwe harde schijf en deze in gebruik wilt nemen, moet je toch alle arrays terug aanmaken.

Wanneer je testen uitvoert, moet je de swap-array zelf terug aanmaken:

```
# mdadm --zero-superblock /dev/sda1
# mdadm --zero-superblock /dev/sdb1
# mdadm --create /dev/md1 -n 2 -l 0 /dev/sda1 /dev/sda2
# mkswap /dev/md1
# swapon /dev/md1
```

## 5.3 Een nieuwe harde schijf erbij.

Zoals reeds vermeld is het het best dat het twee identieke schijven zijn.

Is dit niet mogelijk, dan moet je dezelfde partities aanmaken, ze moeten dezelfde **blocknummers** hebben. Indien dit niet mogelijk is, maak ze dan iets groter dan de goeie schijf.

De methode van aanmaken vind je hierboven terug.

## 5.4 Het superblock van een array stemt niet overeen.

Je schakelt dat specifiek device uit.

```
# mdadm -S /dev/md3
```

Zet van de foute partitie de superblock op 0 met *-zero-superblock*.

```
# mdadm --zero-superblock /dev/sda3
```

Nu moet je de array wel terug aanmaken:

```
# mdadm -C /dev/md3 -n 2 -l 1 /dev/sda3 /dev/sdb3
```

Let op het syncen!

**Hopelijks is het filesystem nog in orde ??????**

## 6 Aanmaken van raid met 1 harde schijf en later de 2de toevoegen.

In de plaats waar de 'missing' harde schijf moet komen, zet je **missing**.

```
# mdadm --create /dev/md1 -n 2 -l 0 /dev/sda1 missing
# mdadm --create /dev/md2 -n 2 -l 1 /dev/sda2 missing
# mdadm --create /dev/md3 -n 2 -l 1 /dev/sda3 missing
```

Je installeert het systeem (*/etc/fstab*, chrooten en */etc/lilo.conf* niet vergeten).

In */etc/lilo.conf* zet je maar 1 schijf:

```
# boot from first raid-blockdevice
boot=/dev/md2
# root is on first raid-blockdevice
root=/dev/md2
# This writes the boot signatures to either disk
raid-extra-boot=/dev/sda,/dev/sdb

...
```

Voer *lilo -t lilo* uit! Je zal wel een foutmelding krijgen dat *sdb* er niet is.

Je koppelt de cdrom af en voegt de 'missing' harde schijf erbij.

Opstarten vanop de 1ste harde schijf.

Gelijke partities aanmaken:

```
sfdisk -d /dev/sda | sfdisk /dev/sdb
```

Rebooten!

De raid-array vervolledigen:

```
# mdadm /dev/md1 -a /dev/sdb1
# mdadm /dev/md2 -a /dev/sdb2
# mdadm /dev/md3 -a /dev/sdb3
```

Bekijk het syncen ...

## 7 Een mdadm config file aanmaken.

Dit hoeft eigenlijk niet, maar het is interessante informatie bij debugging.

```
# echo "DEVICE /dev/sda /dev/sdb" > /etc/mdadm/mdadm.conf
# mdadm --brief --detail --verbose /dev/md1 >> /etc/mdadm/mdadm.conf
# mdadm --brief --detail --verbose /dev/md2 >> /etc/mdadm/mdadm.conf
# mdadm --brief --detail --verbose /dev/md3 >> /etc/mdadm/mdadm.conf
```

## 8 Ondervonden problemen.

- Met de 2.6-kernels ging het niet om te syncen na een fout met een harde schijf. Het systeem hing langzaamaan onder, je kon het mooi volgen. Enkel een ctr-alt-del deed het systeem opnieuw starten.  
Het systeem zelf wou dan niet meer opstarten, enkel met Knoppix kan je de raid-arrays terug herstellen. Dan het systeem opnieuw booten.  
Het vreemde is dat op het net hierover geen enkele thread is terug te vinden???  
Dan maar de 2.4.28 kernel proberen ...  
Hiermee lukt het wel! jihaa