OSL
UNIX Pfadfinder

Root–Spiegelung
mit dem
Solaris Volume Manager

V1.03

Copyright und Handelsmarken

Beschränkungen

Versionen dieses Dokumentes

Version	Datum	Author	e–mail	Inhalt / Änderungen
1.00	05.11.2003	HO	pathfinder@osl–it.de	Erste Fassung
1.01	16.12.2003	MB	pathfinder@osl–it.de	Ergänzungen, Formatierung
1.02	14.01.2004	HO	pathfinder@osl–it.de	Formatierung
1.03	17.06.2004	MB	pathfinder@osl–it.de	Copyright 2004, md_check–Prozedur

Inhaltsverzeichnis

1.Einführung

Dieser Artikel stellt sich zum Ziel, das Anlegen eines Mirrors mit dem Solaris Volume Manager rezeptartig nachvollziehbar darzustellen. Einzelne Schritte sind in chronologischer Reihenfolge aufgeführt.

Stellen Sie bitte sicher, dass folgende Voraussetzungen erfüllt werden.

- mindestens zwei geometrisch identische Festplatten
- ein freies Slice ohne Filesystem (auf beiden Platten) zum Anlegen der Replicas

Folgende Konfiguration soll gespiegelt werden.

```
/dev/dsk/c0t0d0c0 /dev/rdsk/c0t0d0c0 / ufs 1 no –
/dev/dsk/c0t0d0c1 /dev/rdsk/c0t0d0c1 /usr ufs 1 no –
/dev/dsk/c0t0d0c2 /dev/rdsk/c0t0d0c2 /opt ufs 1 no –
```

Begrifflichkeiten:

untergeordneter Spiegel (Submirror): ist ein RAID 0 Gerät (Concat/Stripe), welches in erster Instanz, den Zugriff auf die Slices der Geräte abstrahiert.

Spiegel (Mirror): als Spiegel wird das Gerät bezeichnet, welches als RAID 1 Gerät bis zu 3 untergeordnete Spiegel beschreiben kann.

1.1.Replicas

Nachfolgend werden insgesamt 4 Replicas, davon die ersten zwei auf Slice 3 der ersten Festplatte und zwei weitere auf der zweiten Festplatte angelegt.

```
# metadb –a –f –c 2 c0t0d0s3
# metadb –a –c 2 c0t2d0s3
```

Hinweis: Verteilen Sie ihre Replicas auf verschiedenen Platten. Je mehr Platten, desto höher ist die Wahrscheinlichkeit das bei einem Ausfall genug Replicas für einen automatischen Systemstart vorhanden sind. Ist dies nicht der Fall, d.h. es wurden mehr (oder genau die Hälfte der Replicas) zerstört, wird beim Systemstart ein manuelles Eingreifen notwendig. Bei zwei Spiegeln wäre ein asymmetrisches Verteilen der Replicas überlegenswert, da dies die Wahrscheinlichkeit für ein manuelles Eingreifen beim Systemstart halbieren würde.

1.2.VTOC duplizieren

Sollte es notwendig sein, eine zweite Festplatte mit identischen Slices zu erzeugen, empfiehlt sich folgendes Kommando.

```
# prtvtoc /dev/rdsk/c0t0d0s0 | fmthard –s –
/dev/rdsk/c0t2d0s0
```

Mit diesem Befehl wird das VTOC der ersten Platte in die zweite geschrieben.

1.3.Spiegel initialisieren

Nachfolgend werden die Spiegel erzeugt. Hier im Beispiel wird einem Submirror, bestehend aus einem Slice genau ein Slice hinzugefügt.

Hinweis: Bei den Submirrors dxx kann es sich auch um Stripes oder Concats handeln.

```
metainit -f d10 1 1 c0t0d0s0
metainit -f d20 1 1 c0t2d0s0
metainit -f d11 1 1 c0t0d0s1
metainit -f d21 1 1 c0t2d0s1
metainit -f d12 1 1 c0t0d0s2
metainit -f d22 1 1 c0t2d0s2
```

metainit d10 2 1 c0t0d0s0 c0t0d0s2 definiert ein Concat als d10 bzw.
metainit d10 1 2 c0t0d0s0 c0t0d0s2 steht für ein Stripe Volume.

Hinweis: Wählen Sie ein eingängiges Schema bei der Vergabe der Namen für Submirror. Im obigen Beispiel sind die Platten mit *dxy* bezeichnet, wobei *x* für die Platte (1 oder 2) steht und *y* das Slice kennzeichnet.

1.4.Submirrors einfügen

Nachfolgend werden die Submirrors *dxy* dem Mirror *dz* zugeordnet. Die Geräte *dz* bilden die höchste Ebene der Abstraktion. Sie stellen die Dateisysteme bereit, welche sobald montiert, transparentes Schreiben auf die physikalischen Platten erlauben.

Als erstes fügen wir das Wurzelverzeichnis Root (/) hinzu.

```
# metainit d0 -m d10
# metaroot d0                 # Änderungen in /etc/vfstab
                              automatisch (Solaris 9)
# lockfs -fa
# reboot
...
# metattach d0 d20            # zweiten Spiegel einhängen
```

Bei /usr, /var Filesysteme, bei denen kein aushängen im laufenden Betrieb möglich ist, wird wie folgt verfahren:

```
# metainit d1 -m d11
# vi /etc/vfstab anpassen
Aus Zeile
/dev/dsk/c0t0d0c1 /dev/rdsk/c0t0d0c1  /usr ufs 1 no -
wird
/dev/md/dsk/d1 /dev/md/rdsk/d1        /usr ufs 1 no -
# reboot
...
# metattach d1 d12
```

Alle anderen montierbaren Filesysteme werden wie folgt gespiegelt:

Wichtig: Bei Spiegelung des Root Dateisystems muss das System neu gestartet werden, bevor der zweite Submirror eingehängt werden kann.

```
# metainit d2 -m d12
# umount /opt
# vi /etc/vfstab anpassen
Aus Zeile
/dev/dsk/c0t0d0c2 /dev/rdsk/c0t0d0c2 /opt ufs 1 no -
wird
/dev/md/dsk/d2 /dev/md/rdsk/d2 /opt ufs 1 no -
# mount /opt
...
# metattach d2 d22
```

Hinweis: Das *metaroot* Kommando (Spiegelung von root (/)) fügt */etc/vfstab* und */etc/systems* entsprechende Änderungen hinzu. Diese Dateien müssen somit nicht angepasst werden. (Solaris Version 9)

1.5.Submirror entfernen

Mit den Kommandos *metadetach* werden die Submirrors vom Mirror (in unserem Beispiel die Geräte d1, d2, d3) entfernt. Bevor allerdings der letzte verbleibende Submirror entfernt werden kann, muss der Mirror abmontiert werden, erst dann kann mit *metaclear* selbiger entfernt werden.

```
entfernen des Submirrors d13
# metadetach d3 d13
entfernen des Mirrors d3
# metaclear d3
```

Wichtig: Zum Entfernen der Mirror müssen diese abmontiert sein, d.h. etwaige Änderungen in der */etc/vfstab* müssen rückgängig gemacht, und das System neu gestartet werden. Änderungen für das Wurzelverzeichnis root (/) werden mit dem Kommando *metaroot /dev/dsk/c0t0d0s0* rückgängig gemacht.

Sobald die oben gezeigten Spiegelungen vorgenommen wurden, sollten die Boot Variablen angepasst werden. Je nach Vorliebe kann dies am Openboot Prompt oder in einem Terminalfenster (mit eeprom) geschehen. Dieses Anpassen ist sinnvoll, soll doch erstens bei Ausfall einer Platte automatisch von der anderen gestartet werden und zweitens den Platten ihrer Funktion (Bootdisk–Mirrordisk, Master–Slave) entsprechend, Namen gegeben werden.

2.Openboot Prom (OBP)

Zuerst aber ist es notwendig die Geräteadressen der Submirror zu notieren.

```
# ls -l /dev/dsk/c0t0d0s0
lrwxrwxrwx 1 root root 55 Mar 5 12:54
/dev/rdsk/c0t0d0s0 -> \
../../devices/pci@1f,0/ide@d/dad@0,0:a
# ls -l lrwxrwxrwx 1 root root 55 Mar 5 12:54
/dev/rdsk/c0t2d0s0 -> \
../../devices/pci@1f,0/ide@d/dad@2,0:a
```

Der jeweils unterstrichene Teil kennzeichnet die Geräteadresse.

Wechseln Sie nun zum Openboot Prompt.

```
Zum Openbootprompt wechseln
# init 0            # an Kommando Zeile
...
ok                  # OBP Prompt
```

Im OBP kann nun ein entsprechender Alias auf die oben erzeugten Submirror gesetzt werden. Im Beispiel ist das Gerät *c0t2d0s0* die Masterkopie. Folgende Kommandos schreiben die Aliase *bootdisk* und *mirrordisk* in den nicht flüchtigen RAM des OBP. Einträge in diesen RAM müssen mit einem *nvstore* abgeschlossen werden, da nur dann der RAM in einen konsistenten Zustand gesetzt wird. Aliase lassen sich mit *nvunalias* löschen. Gelöschte Alias Einträge verschwinden erst nach einem *reset–all* am Openboot Prompt.

```
ok nvalias bootdisk /pci@1f,0/ide@d/dad@0,0:a
ok nvalias mirrordisk /pci@1f,0/ide@d/dad@2,0:a
ok nvstore
```

Hinweis: Sollten IDE Geräte nicht mit diesem Alias starten können, ersetzen Sie die Zeichenkette *dad* (oder *sd* im Falle von SCSI Platten) durch *disk*. Um herauszufinden wie OBP die Geräte anspricht, lassen sie sich diese mit dem Befehl *devalias* anzeigen. Dieser Befehl zeigt alle momentan definierten Aliase an. In der Standardeinstellung des OBP existieren bereits Einträge für diverse Geräte wie z.B. Festplatten, CD–Laufwerke etc.

Nachfolgend wird die OBP Umgebungsvariablen *boot–device* mit den Aliasen (Geräten) *bootdisk* und *mirrordisk* gesetzt. Diese Variable bestimmt die Bootgeräte und deren Reihenfolge im Bootprozess. Umgebungsvariablen werden mit den Befehlen *printenv* angezeigt und mit *setenv* gesetzt.

```
ok printenv boot-device
boot-device =      disk net
ok setenv boot-device bootdisk mirrordisk
```

Mit dem Solaris Werkzeug *eeprom* steht eine weitere Möglichkeit die Bootvariablen zu setzen zur Verfügung. So ändert der Befehl

```
# eeprom boot-device='mirrordisk bootdisk'
```

die Bootreihenfolge äquivalent zum oben gezeigten *setenv* am Openboot Prompt. Ebenso ist es möglich die Gerätealiase auf der Kommandozeile einzugeben.

```
# eeprom nvramrc='devalias bootdisk
/pci@1f,0/ide@d/dad@0,0:a devalias mirrordisk
/pci@1f,0/ide@d/dad@2,0:a'
```

Hinweis: Nehmen sie sich die Zeit und ändern Sie die Namen der Festplatten, anstatt bereits existierende Aliase auf ihre Geräte umzubiegen. Dies hat zum einen den Vorteil, dass Sie ihre momentane Bootreihenfolge und die Geräteadressen auf einen Blick von *eeprom* aus sehen können. (eeprom | grep boot–device und eeprom | grep nvramrc). Hartcodierte Standardeinträge für ihre Geräte wie (disk1 etc.) sind von *eeprom* aus nicht sichtbar, Sie können also nicht sicher sein, ob der *disk6* Alias auch wirklich auf dieses Gerät zeigt.

3.Troubleshooting

3.1.Ausfall einer Festplatte

Symptom: System bootet im Single User Mode und zeigt untenstehende Meldung

```
metainit: hostname: stale databases


Insufficient metadevice database replicas located.


Use metadb to delete databases which are broken.
Ignore any ^Read-only file system~ error messages.
Reboot the system when finished to reload metadevice
database.
After reboot, repair any broken database replicas which
were deleted.
```

Lösung: Entfernen Sie die nicht erreichbaren Replicas

Achtung: Die im Folgenden benutzten Gerätenamen sind nicht mit der Ihnen vorliegenden Konfiguration identisch. Bitte passen sie die Gerätenamen den Ihrigen an.

Melden sie sich mit ihrem root Passwort an und finden sie heraus welche Festplatte bzw. Replicas betroffen sind.

```
# metadb -i
        flags     first blk          block count
M       p         16        unknown  /dev/dsk/c0t0d0s3
M       p         8208      unknown  /dev/dsk/c0t0d0s3
...
```

Im nächsten Schritt müssen die Replicas entfernt und das System neu gestartet werden. Nach dem Neustart können die Fehlermeldungen ignoriert und mit CTRL–D das Booten in den Multi–User Betrieb fortgesetzt werden.

```
# metadb -d -f c0t0d0s3
# reboot
```

Wichtig: Ihr System arbeitet momentan mit nur einer Platte. Es ist demzufolge außerordentlich wichtig die zweite Platte wiederherzustellen.

3.2.Platte wiederherstellen

Zum Herstellen einer 1 zu 1 Kopie duplizieren Sie das VTOC wie oben beschrieben, stellen Sie dabei sicher das ***nicht*** ein eventuell defektes VTOC kopiert wird.

D.h. wenn Platte mit Slice */dev/rdsk/c0t0d0s0* defekt ist, und */dev/rdsk/c0t2d0s0* ein intaktes VTOC besitzt, wird mit dem Kommando:

```
# prtvtoc /dev/rdsk/c0t2d0s0 | fmthard -s -
/dev/rdsk/c0t0d0s0
```

das defekte VTOC wiederhergestellt.

Im nächsten Schritt sind die Datenbank Replicas wiederherzustellen.

```
# metadb -a -f -c 2 c0t0d0s3
```

Verifizieren sie die erfolgreiche Wiederherstellung mit

```
# metadb -i
```

Im Folgenden werden die Spiegel wieder dem Metageräten hinzugefügt.

```
# metareplace -e d1 c0t0d0s0
# metareplace -e d2 c0t0d0s1
# metareplace -e d3 c0t0d0s2
```

Nach diesem Schritt können Sie das Synchronisieren der Platten mit

```
# metastat
```

verfolgen. Abhängig von Plattengröße und Leistung kann dieser Vorgang einige Zeit in Anspruch nehmen.

Sollten sich die Volumes nicht auf diese Weise wiederherstellen lassen, empfiehlt sich ein erneutes Aufsetzen. Hierbei müssten die Einträge in der */etc/vfstab* entfernt, das System neu gestartet und die Spiegel wie oben beschrieben konfiguriert werden.

Wichtig: Defekte Replicas werden, wenn sie nicht vom Solaris Volume Manager repariert werden können, ignoriert und müssen manuell instand gesetzt werden.

Hinweis: Einem Verlust aller Replicas folgt unweigerlich der Verlust aller Solaris Volume Manager Volumes. Es ist also wichtig genug Replicas anzulegen und diese ausfallsicher zu verteilen.

3.3.Oft gestellte Fragen

Frage: **Ich will Mirror bzw. Submirror entfernen ? Es sind keine Metagerät montiert, trotzdem melden *metadetach* und *metaclear* Device busy oder ähnliches. Wie kann ich sie trotzdem entfernen?**

Lösung: Stellen Sie sicher, dass keines dieser Geräte montiert ist und löschen Sie die Spiegel mit dem f Schalter.

```
# metaclear -f dx.
```

Frage: **Wie kann ich mir im OBP alle Umbebungsvariablen und ihren momentanen Wert anzeigen lassen?**

Lösung:

```
ok printenv
Variable Name                    Value                Default Value

test-args
...
boot-device
...
```

Frage: **Ich möchte gern alle definierten Spiegel und die dazugehörigen untergeordneten Spiegel sehen. Wie mache ich das ?**

Lösung: Die datei */etc/lvm/md.tab* enthält die Konfigurationsdaten Ihrer Spiegel. Mit

```
# metastat -p
```

erhalten Sie eine Übersicht.

Frage: **Wie werde ich darauf aufmerksam, daß Platten ausgefallen sind?**

Antwort: Dies ist eines der Hauptprobleme bei hostbasierter Spiegelung. Eindeutig liegt dies im Aufgabenbereich des Unix–Sytemverwalters, und: Das System wird in der Regel häufiger ausfallen als ein Enterprise RAID–System. Jeder dieser Ausfälle kann – abgesehen von Plattenausfällen – auch Probleme im Umfeld der Spiegelung nach sich ziehen.
Einige Anregungen dazu finden sich im "Solaris Volume Manager – Administration Guide" .
Wir haben ein simples Script »md_check«, das bei etlichen Fehlersituationen eine Nachricht z. B. über syslog nach /var/adm/messages schreibt (bitte testen, auf geeignete syslog–Konfiguration achten und ggf. anpassen). Das Script testet – ohne weitere Argumente aufgerufen – die internen Metadisks. Man kann auch Disksets prüfen (einfache Cluster). Der Aufruf erfolgt z. B. zur Prüfung der Root–Spiegelung und eines externen Disksets »jumbo« per crontab–Eintrag wie folgt:

```
05 09 * * * /usr/local/bin/md_check > /dev/null 2>&1
10 09 * * * /usr/local/bin/md_check -s jumbo  > /dev/null 2>&1
```

Nur sofern ein Problem entdeckt wird, erscheint dann in /var/adm/messages folgende Meldung:

```
Jun 08 09:05:00 pumuckl root: [ID 302631 daemon.notice] WARNING (md_check):
There are problems with local md configuration!
```