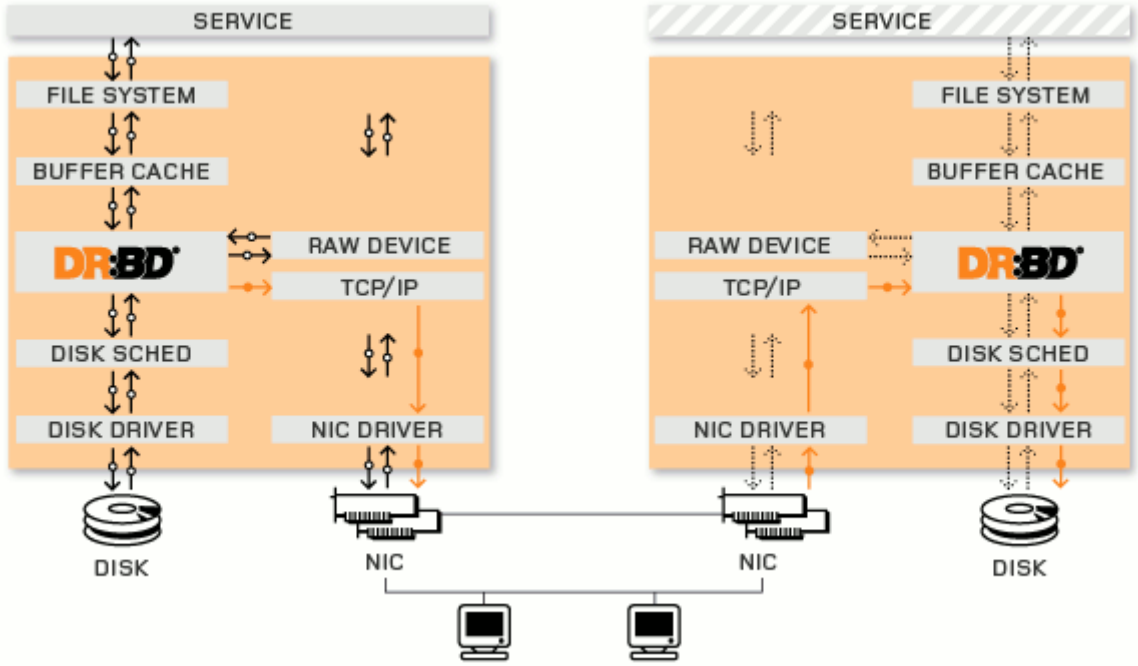
Preface

本文採用 Open Source 套件 DRBD(Distributed Replicated Block Device)及 Heartbeat 作為兩台 Linux 伺服器高可用性叢集架構(HA Cluster)。DRBD 可在網路環境下，進行資料鏡像(Mirror)的同步，不同於檔案的複製，而是利用區塊(Block)來進行資料的搬移及交換。DRBD 也可將之視作為以網路為基礎的 RAID-1。

下圖為 DRBD HA Cluster 架構流程圖。



LAB Environment

OS: CentOS 6.3 (2.6.32-279.el6.x86_64)

Primary Server (DRBD-1)	eth0	172.20.144.198 (Mgmt)
	eth1	1.1.1.1 (Mirror)

	Mirror HDD	/dev/sdb
Secondary Server (DRBD-2)	eth0 eth1 Mirror HDD	172.20.144.199 (Mgmt) 1.1.1.2 (Mirror) /dev/sdb
Cluster IP	eth0:0	172.20.144.200
DRBD Resource Name	r0	

1. 安裝必要套件(YUM 自動安裝)

I. 新增以下 3 個 YUM Repository。若連結失效請自行至該網站找最新連結。

II. 若無法使用 YUM 來自動安裝，可參考下一則手動安裝參考。

RPMforge
rpm -Uvh
http://pkgs.repoforge.org/rpmforge-release/rpmforge-release-0.5.2-2.el6.rf.x86_64.rpm

ELRepo
rpm -Uvh http://elrepo.org/elrepo-release-6-4.el6.elrepo.noarch.rpm

EPEL
rpm -Uvh http://dl.fedoraproject.org/pub/epel/6/x86_64/epel-release-6-7.noarch.rpm

III. 安裝系統相關及編譯套件

```
yum -y install gcc policycoreutils-python setroubleshoot system-config-services
```

IV. 安裝 DRBD 相關套件

```
yum -y install drbd83-utils kmod-drbd83
```

V. 安裝 Heartbeat 相關套件

```
yum -y install heartbeat heartbeat-devel heart-libs
```

2. 安裝必要套件(無法連至 Internet，需手動安裝)

I. 於 CentOS 安裝時，選擇 Basic Server 安裝。

II. 設定 YUM CentOS Disc 套件安裝

i. 使用 YUM 可自動安裝相依性套件，省去手動安裝麻煩。

ii. 放入光碟，並掛載於 /mnt。

iii. 修改 /etc/yum.repos.d/CentOS-Media.repo，刪除原先內容，並加入以下資料，下列內容請參考 CentOS 6.3 光碟目錄內的 .discinfo。

```
[InstallMedia]
name=6.3
baseurl=file:///mnt/
mediaid=1341569670.539525
metadata_expire=-1
gpgcheck=0
cost=500
```

III. 準備 DRBD 安裝相關套件

drbd83-utils-8.3.13-1.el6.elrepo.x86_64.rpm

kmod-drbd83-8.3.13-2.el6_3.elrepo.x86_64.rpm

IV. 安裝 DRBD 套件

```
# yum -y localinstall drbd83-utils-8.3.13-1.el6.elrepo.x86_64.rpm
# yum -y localinstall kmod-drbd83-8.3.13-2.el6_3.elrepo.x86_64.rpm
```

V. 安裝 Heartbeat 套件

```
# yum -y localinstall heartbeat-3.0.4-1.el6.x86_64.rpm
heartbeat-devel-3.0.4-1.el6.x86_64.rpm heartbeat-libs-3.0.4-1.el6.x86_64.rpm
```

3. 伺服器主機基礎設定 (兩台主機都需設定)

I. 主機名稱(Hostname)定義

- Primary Server: DRBD-1
- Secondary Server: DRBD-2

II. 防火牆及 SELinux 設定

III. 修改 /etc/hosts，加入兩台主機名稱及 IP。

```
[root@DRBD-1 ~]# vim /etc/hosts
172.20.144.198    DRBD-1
172.20.144.199    DRBD-2
```

IV. IP 配置，包含管理 IP(172.20.144.x)及鏡像 IP(1.1.1.x)

V. 配置鏡像磁碟，本例使用裝置 /dev/sdb。

```
[root@DRBD-1 ~]# fdisk /dev/sdb

Command (m for help): n
```

```
Command action
   e   extended
   p   primary partition (1-4)
p
Partition number (1-4): 1
First cylinder (1-6527, default 1):
Using default value 1
Last cylinder, +cylinders or +size{K,M,G} (1-6527, default 6527):
Using default value 6527

Command (m for help): p

Disk /dev/sdb: 53.7 GB, 53687091200 bytes
255 heads, 63 sectors/track, 6527 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes
Sector size (logical/physical): 512 bytes / 512 bytes
I/O size (minimum/optimal): 512 bytes / 512 bytes
Disk identifier: 0x0009d6ac

   Device Boot      Start         End      Blocks   Id  System
/dev/sdb1               1        6527    52428096   83  Linux

Command (m for help): w
The partition table has been altered!

Calling ioctl() to re-read partition table.
Syncing disks.
```

VI. 於根目錄建立 DRBD 預備掛載資料夾

```
[root@DRBD-1 ~]# mkdir /DRBD
```

4. 修改 drbd 設定檔 (兩台主機都需設定)

I. 修改 DRBD 設定檔 (簡易設定)

```
[root@DRBD-1 ~]# vim /etc/drbd.d/global_common.conf
global {
    usage-count no;
}
common {
    syncer { rate 1000M; }
}
```

II. 修改 DRBD 資源設定檔 (簡易設定)

```
[root@DRBD-1 ~]# vim /etc/drbd.d/r0.res
resource r0 {
    protocol c;
    on DRBD-1 {
        device      /dev/drbd1;
        disk        /dev/sdb1;
        address     1.1.1.1:7789;
        meta-disk internal;
    }
    on DRBD-2 {
        device      /dev/drbd1;
        disk        /dev/sdb1;
        address     1.1.1.2:7789;
        meta-disk internal;
    }
}
```

5. 啟用裝置資源(Resource) (兩台主機都需執行)

I. 建立裝置 Metadata

```
[root@DRBD-1 /]# drbdadm create-md r0
Writing meta data...
initializing activity log
NOT initialized bitmap
New drbd meta data block successfully created.
```

II. 啟動 DRBD 服務，並設定開機自動啟動。

```
[root@DRBD-1 /]# service drbd start
Starting DRBD resources: [ d(r0) s(r0) n(r0) ].
[root@DRBD-1 /]# chkconfig drbd on
```

III. 檢查 DRBD 服務目前狀態，由於尚未指派 Primary，所以目前狀態都是 Secondary。

```
[root@DRBD-1 ~]# cat /proc/drbd
version: 8.3.13 (api:88/proto:86-96)
GIT-hash: 83ca112086600faacab2f157bc5a9324f7bd7f77 build by dag@Build64R6,
2012-09-04 12:06:10

 1: cs:Connected ro:Secondary/Secondary ds:Inconsistent/Inconsistent C r-----
    ns:0 nr:0 dw:0 dr:0 al:0 bm:0 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:10482024
```

cs: 主從狀態。也可以用指令”drbdadm role r0”查詢。

ds: 磁碟狀態

IV. 設定 Primary 主機，並開始同步兩台主機磁碟資料。(僅需 Primary Server 執行)

```
[root@DRBD-1 ~]# drbdadm -- --overwrite-data-of-peer primary r0
```

V. 查詢 Primary Server 同步狀 (同步進行中)

```
[root@DRBD-1 ~]# service drbd status
drbd driver loaded OK; device status:
version: 8.3.13 (api:88/proto:86-96)
GIT-hash: 83ca112086600faacab2f157bc5a9324f7bd7f77 build by dag@Build64R6,
2012-09-04 12:06:10
m:res   cs         ro                   ds                         p  mounted  fstype
...     sync'ed:   2.5%                 (9984/10236)M
```

```
1:r0     SyncSource   Primary/Secondary   UpToDate/Inconsistent   C
```

VI. 查詢 Secondary Server 同步狀態 (同步進行中)

```
[root@DRBD-2 /]# cat /proc/drbd
version: 8.3.13 (api:88/proto:86-96)
GIT-hash: 83ca112086600faacab2f157bc5a9324f7bd7f77 build by dag@Build64R6,
2012-09-04 12:06:10

 1: cs:SyncTarget ro:Secondary/Primary ds:Inconsistent/UpToDate C r-----
    ns:0 nr:482048 dw:473856 dr:0 al:0 bm:28 lo:65 pe:7370 ua:64 ap:0 ep:1 wo:b
oos:10008168
        [>.....................] sync'ed:    4.6% (9772/10236)M
        finish: 0:05:37 speed: 29,616 (29,616) want: 1,024,000 K/sec
```

VII. 查詢 Primary Server 同步狀態 (同步完成)

```
[root@DRBD-1 ~]# service drbd status
drbd driver loaded OK; device status:
version: 8.3.13 (api:88/proto:86-96)
GIT-hash: 83ca112086600faacab2f157bc5a9324f7bd7f77 build by dag@Build64R6,
2012-09-04 12:06:10
m:res   cs           ro                   ds                   p   mounted   fstype
1:r0    Connected    Primary/Secondary    UpToDate/UpToDate    C
```

VIII. 查詢 Secondary Server 同步狀態 (同步完成)

```
[root@DRBD-2 /]# cat /proc/drbd
version: 8.3.13 (api:88/proto:86-96)
GIT-hash: 83ca112086600faacab2f157bc5a9324f7bd7f77 build by dag@Build64R6,
2012-09-04 12:06:10

 1: cs:Connected ro:Secondary/Primary ds:UpToDate/UpToDate C r-----
    ns:0 nr:10482024 dw:10482024 dr:0 al:0 bm:640 lo:0 pe:0 ua:0 ap:0 ep:1 wo:b oos:0
```

6. 開始使用 DRBD

I. 為/dev/drbd1 格式化，並建立 EXT4 檔案系統。 (僅需 Primary Server 執行)

```
[root@DRBD-1 ~]# mkfs.ext4 /dev/drbd1
mke2fs 1.41.12 (17-May-2010)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
655360 inodes, 2620506 blocks
131025 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=2684354560
80 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks:
        32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632

Writing inode tables: done
Creating journal (32768 blocks): done
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 31 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
```

II. 將 /dev/drbd1 掛載至 /DRBD ，並檢查掛載情形。 (僅需 Primary Server 執行)

```
[root@DRBD-1 ~]# mount /dev/drbd1 /DRBD
[root@DRBD-1 ~]# df
Filesystem            1K-blocks      Used Available Use% Mounted on
/dev/sda1             16512936    2656832  13017296  17% /
tmpfs                   510320          0    510320   0% /dev/shm
/dev/drbd1            10317472     154100   9639272   2% /DRBD
```

7. DRBD 斷線測試

I. 將 Primary Server 網路埠 eth1 的網路線拔掉或下指令斷線

```
[root@DRBD-1 ~]# ifdown eth1
```

II. 這時檢查 Primary Server 的 drbd 狀態，發現 Secondary 變成 Unknown。

```
[root@DRBD-1 ~]# drbdadm role r0
Primary/Unknown
```

III. 而 Secondary Server 的 drbd 狀態，Primary 變成 Unknown。

```
[root@DRBD-2 ~]# drbdadm role r0
Unknown /Primary
```

IV. 恢復網路連線，並執行 drbd reconnect 重新連接指令。

```
[root@DRBD-1 ~]# ifup eth1
[root@DRBD-1 ~]# drbdadm connect r0
[root@DRBD-1 ~]# drbdadm role r0
Primary/Secondary
```

8. DRBD 測試資料是否同步鏡像寫入

I. 在 Primary Server 的 DRBD 目錄底下新增一個名為 jtest.iso 檔案。完成後，卸載 /DRBD 裝置。

```
[root@DRBD-1 ~]# dd if=/dev/zero of=/DRBD/jtest.iso bs=1M count=1024
[root@DRBD-1 ~]# umount /DRBD
```

II. 將 Primary Server 的 drbd 角色轉換為 Secondary Server。

```
[root@DRBD-1 ~]# drbdadm secondary r0
[root@DRBD-1 ~]# drbdadm role r0
Secondary/Secondary
```

III. 將 Secondary Server 的 drbd 角色轉換為 Primary Server。

```
[root@DRBD-2 ~]# drbdadm primary r0
[root@DRBD-2 ~]# drbdadm role r0
Primary/Secondary
[root@DRBD-1 ~]# drbdadm role r0
Secondary/Primary
```

IV. 在 Secondary Server 掛載 /DRBD，並檢查檔案是否正確被同步寫入。

```
[root@DRBD-2 /]# mount /dev/drbd1 /DRBD
[root@DRBD-2 /]# ll /DRBD
total 1112848
-rw-r--r--. 1 root root 1073741824 Nov 18 19:16 jtest.iso
```

9. 手動腦裂(Split brain)復原

I. 腦裂(Split brain)是指在某種情況下，造成 DRBD 兩個節點斷開了連接，都以 Primary 的身份來運行。當 DRBD 某 primary 節點連接對方節點準備發送訊息的時候如果發現對方也是 primary 狀態，那麼會立刻自行斷開連接，並認定當前已經發生 split brain 了，這時候他會在系統日志中記錄以下信息： “Split-Brain detected,dropping connection!” 當發生 split brain 之後，如果查看連接狀態，其中至少會有一個是 StandAlone 狀態，另外一個可能也是 StandAlone（如果是同時發現 split brain 狀

態)，也有可能是 WFConnection 的狀態。

II. 在手動復原前，需先確認好哪一台 Server 是正確的 Primary。確定後，到另一台 Server，切換成 Secondary 角色，並放棄 Secondary 的資源資料。

```
[root@DRBD-2 ~]# drbdadm secondary r0
[root@DRBD-2 ~]# drbdadm disconnect r0
[root@DRBD-2 ~]# drbdadm -- --discard-my-data connect r0
```

III. 若 Primary Server 當前狀態為 WFConnection，則會開始自動重新同步。若狀態依然為 StandAlone，則需手動下連接的指令。

```
[root@DRBD-1 ~]# drbdadm connect r0
```

10.Heartbeat

I. 在預設只有安裝 DRBD 情況下，Primary/Secondary 主機只能靠手動去切換，很不方便也不即時。下列將透過 Heartbeat 配置達到兩台主機間自動切換。

II. 首先，新增 Heartbeat 服務於系統中。(兩台主機都要設定)

```
[root@DRBD-1 ~]# chkconfig --add heartbeat
[root@DRBD-1 ~]# chkconfig heartbeat on
```

III. Heartbeat 有三個設定檔，分別為 ha.cf、haresources 及 authkkeys，將此三個範例設定檔複制到 /etc/ha.d/ 目錄底下或手動新增。(兩台主機都需要)

```
[root@DRBD-1 ~]# cp /usr/share/doc/heartbeat-3.0.4/ha.cf    /etc/ha.d/
[root@DRBD-1 ~]# cp /usr/share/doc/heartbeat-3.0.4/haresources    /etc/ha.d/
[root@DRBD-1 ~]# cp /usr/share/doc/heartbeat-3.0.4/authkeys    /etc/ha.d/
```

IV. 編輯 Heartbeat 主要配置檔 ha.cf (兩台主機都需設定)

[root@DRBD-1 ~]# vim /etc/ha.d/ha.cf

```
autojoin none

# Debug messages
debugfile /var/log/ha-debug

# Other messages
logfile /var/log/ha-log

# Syslog 系統日誌
logfacility local0

#  多久時間確認對方是否存活(單位:秒)
keepalive 2

#  多久時間確認對方已完全失去聯繫(單位:秒)
deadtime 10

#  連續多久時間聯繫不上後開始警告提示(單位:秒)
warntime 4

#  主機若重開機，等待網路開啟或其他應用程式執行的時間。(單位:秒)
initdead 60


#  使用 694 Port 來作 Heartbeat 監控

udpport 694


#  採用網路卡 eth1 的 UDP 廣播來發送 heartbeat 訊息

#bcast eth1


#  採用網路卡 eth1 的 UDP 單播  來發送 heartbeat 訊息，IP 為對方 IP 位置。建議採用單
```

播，避免多組 cluster 主機都會看到對方節點(第二台主機則填對方 IP)。

```
ucast eth0 172.20.144.199
ucast eth1 1.1.1.2

# 若 Primay 異常修復完畢，是否需要從 Secondary 自動切換回 Primary。建議設定 off。
auto_failback off

# 節點 1 與節點 2，必須要與 uname -n 指令得到的名稱一致。
node DRBD-1
node DRBD-2

ping 172.20.144.254

respawn hacluster /usr/lib64/heartbeat/ipfail
respawn hacluster /usr/lib64/heartbeat/dopd
apiauth dopd gid=haclient uid=hacluster
```

V. 編輯 Heartbeat 認證資訊檔 authkeys (兩台主機 authkeys 內容需一致)

i. Authkeys 為 Cluster 節點間相互認證的密碼，也就是說主機必須擁有此密碼才可加入該 Cluster 群組。Authkeys 共有三種認証方式，分別為 crc、md5 及 sha1，依安全性等級區分，sha1 最不易破解，其次 md5，最後為 crc。

ii. Authkeys 的權限必須為 600

iii. 兩台主機 authkeys 內容需一致

```
[root@DRBD-1 ~]# (echo -ne "auth 1\n1 sha1 "; echo $RANDOM | openssl sha1 | awk '{print $2}') > /etc/ha.d/authkeys
[root@DRBD-1 ~]# chmod 600 /etc/ha.d/authkeys
[root@DRBD-1 ~]# scp /etc/ha.d/authkeys root@DRBD-2:/etc/ha.d/
```

```
# authkeys 開啟內容如下

auth 1
1 sha1 ccde56f52bc06bcda0918cfa9439f5c91d941de6
```

VI. 編輯 Heartbeat 資源檔 haresources (兩台主機內容需一致)

i. Haresources 每一行代表一個資源組

ii. 資源組的啟動順序是由左至右，關閉順序是由右往左。

iii. Script 的參數是由 :: 來傳遞及分隔。

iv. 每一個資源則以空格間隔

```
[root@DRBD-1 ~]# vim /etc/ha.d/haresources
DRBD-1 172.20.144.200 drbddisk::r0 Filesystem::/dev/drbd1::/DRBD::ext4 ftpd
```

參數說明:	
DRBD-1	指定 Primary 節點
172.20.144.200	Cluster IP 位址
drbddisk::r0	DRBD Resource 名稱
Filesystem::/dev/drbd1::/DRBD::ext4	{Device}::{Mount Point}::{檔案系統型態}
ftpd	欲啟動 daemon 服務

VII. 啟動 Heartbeat 服務 (兩台主機)

```
[root@DRBD-1 ~]# service heartbeat start; ssh drbd-2 service heartbeat start
[root@DRBD-1 ~]# service heartbeat status ; ssh drbd-2 service heartbeat status
heartbeat OK [pid 3352 et al] is running on drbd-1 [drbd-1]...
heartbeat OK [pid 3003 et al] is running on drbd-2 [drbd-2]...
```

VIII. 啟動完 Heartbeat 服務後，確認 Primary 主機是否具備以下狀態。

i. DRBD Disk 掛載於 Primary 主機

ii. Heartbeart 新增一組 Cluster IP

```
[root@DRBD-1 ~]# mount
/dev/drbd1 on /DRBD type ext4 (rw)
[root@DRBD-1 ~]# ifconfig
eth0:0      Link encap:Ethernet  HWaddr 00:50:56:BD:45:CC
            inet addr:172.20.144.200  Bcast:172.20.144.255  Mask:255.255.255.0
            UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
```

11. Heartbeat 各種災害情況演練

I. 以下針對 HA 架構下 Primary 與 Secondary 可能進行切換、服務移轉的情境測試。包括手動切換節點、拔除 Primary 網路線及 Shutdown Primary 主機。

II. HA 服務移轉時間長短，係依據 Heartbeat 設定檔(/etc/ha.d/ha.cf)參數來決定移轉的時間。

III. 關閉 Primary 主機 Heartbeat 服務

```
[root@DRBD-1 ~]# service heartbeat stop
Stopping High-Availability services: Done.
Waiting to allow resource takeover to complete:Done.

[root@DRBD-1 ~]# drbdadm role r0
Secondary/Primary
[root@DRBD-2 ~]# drbdadm role r0
Primary/Secondary
```

IV. 拔除 Primary 對外(eth0)網路線

V. 關閉 Primary 主機

12.Primary / Secondary 主機狀態確認及檢查

I. Primary 主機通常會具備以下條件

i. 執行主要系統端相關服務，例如: FTP、MySQL 等服務。

ii. 網卡被綁定 Cluster IP (eth0:0)

iii. 掛載 DRBD Disk

II. HA 服務移轉時間長短，係依據 Heartbeat 設定檔(/etc/ha.d/ha.cf)參數來決定移轉的時間。

III. 關閉 Primary 主機 Heartbeat 服務

```
[root@DRBD-1 ~]# service heartbeat stop
Stopping High-Availability services: Done.
Waiting to allow resource takeover to complete:Done.

[root@DRBD-1 ~]# drbdadm role r0
Secondary/Primary
[root@DRBD-2 ~]# drbdadm role r0
Primary/Secondary
```

IV. 拔除 Primary 對外(eth0)網路線

V. 關閉 Primary 主機