

# Cluster con Vserver, DRBD e heartbeat

Alberto Cammozzo  
mmzz @ pluto.it

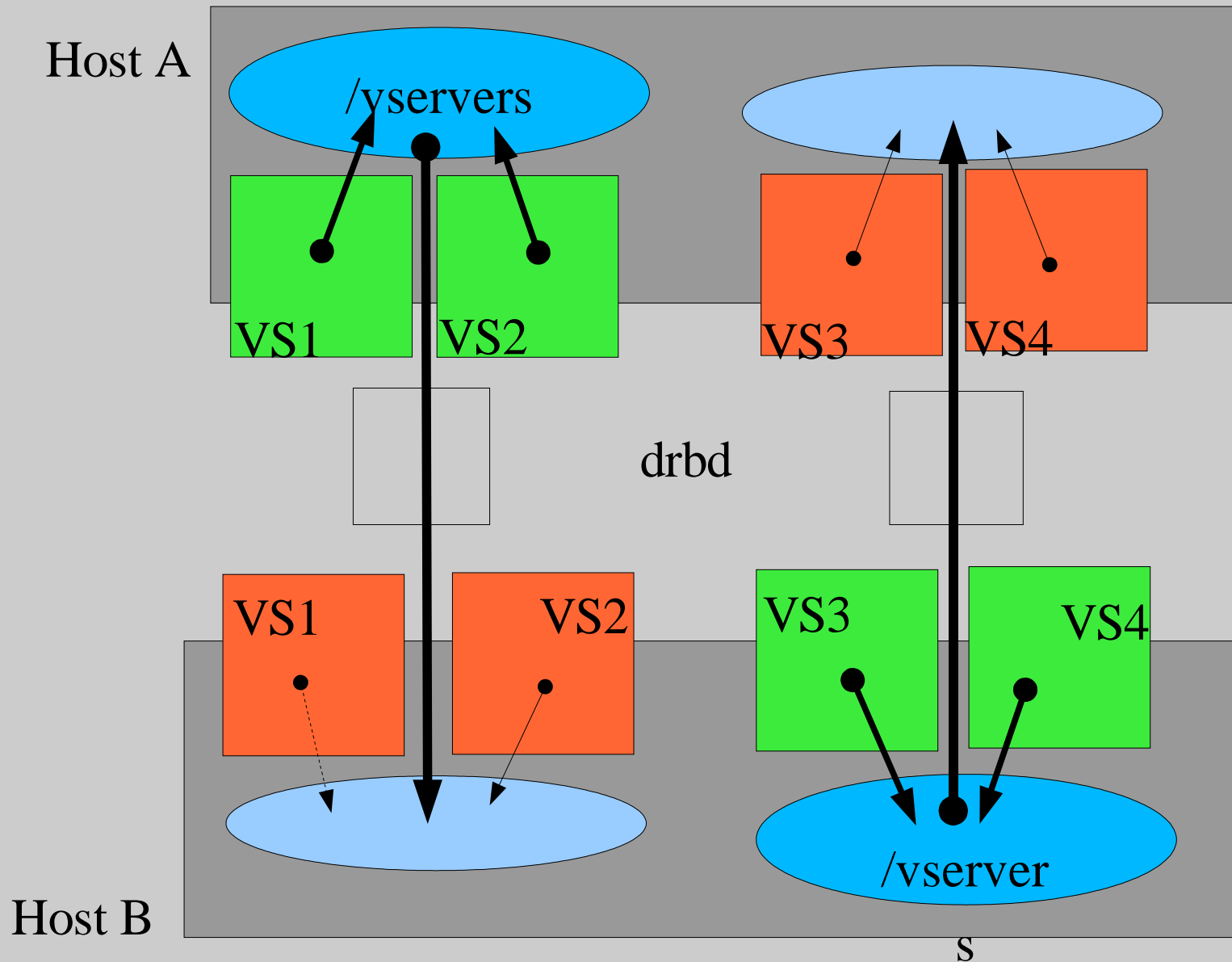
23 novembre 2004

serate a tema PLUTO Padova

# A che serve?

- DRBD:
  - emula uno storage condiviso su storage locale
  - con mirroring (sorta di RAID-1)
- + Heartbeat:
  - Cluster alta affidabilità senza load-balancing
  - 1 server attivo,
  - 1 server hot stand-by.
- Costa poco: il raddoppio dello spazio su disco.

# HA vserver per i poveri (drbd)



# Vserver

- Un solo server hardware
- Un solo kernel
- Diversi vserver indipendenti:
  - processi
  - networking TCP/IP e Sys V IPC
  - filesystem
  - distribution
  - servizi
  - utenti (incluso root)

# DRBD - Distributed Replicated Block Device

- Replica un block device (SCSI, EIDE, ...)
  - da una macchina
  - a una sola altra macchina
  - attraverso una connessione veloce di rete
  - in numerazione IP.
- Ogni device `/dev/nbX` può essere:
  - *primary* (fa il write localmente e sul nodo remoto)
  - *secondary* (fa il write dal *primary*).
- Write: locale e remoto. Read: sempre locale.
- Non si occupa della consistenza del filesystem.

# Heartbeat

- Hardware tra i due nodi:
  - connessione ethernet
  - connessione seriale
  - ...
- Programma heartbeat user-space
- Tool di supporto (*gratuitous arp*)
- Moduli (shell script) per ogni servizio da rendere fault-tolerant sui due nodi
- Un nodo solo per volta può ospitare un dato servizio (no *load-balancing*).

# DRBD + heartbeat

- Drbd ha un modulo heartbeat specifico
- Se il nodo primary si guasta, heartbeat
  - Si occupa di eventuale fsck
  - commuta il secondary in primary
  - fa partire i servizi.
- Se il nodo guasto ritorna attivo,
  - Drbd lo mette secondary, si risincronizza col primary (senza interruzione)
  - Heartbeat lo fa ridiventare primary e fa ripartire i servizi, mettendo in secondary l'altro nodo.

# Configurazione Rete

novanta:~# cat >> /etc/network/interfaces

```
auto lo
iface lo inet loopback
```

```
auto eth1
iface eth1 inet static
  address 147.162.35.90
  netmask 255.255.255.0
  network 147.162.35.0
  broadcast 147.162.35.255
  gateway 147.162.35.254
```

```
auto eth0
iface eth0 inet static
  address 192.168.90.90
  netmask 255.255.255.0
  broadcast 192.169.90.255
```

novantuno:~# cat >> /etc/network/interfaces

```
auto lo
iface lo inet loopback

auto eth1
iface eth1 inet static
  address 147.162.35.91
  netmask 255.255.255.0
  network 147.162.35.0
  broadcast 147.162.35.255
  gateway 147.162.35.254
```

```
auto eth0
iface eth0 inet static
  address 192.168.90.91
  netmask 255.255.255.0
  broadcast 192.169.90.255
```



Cavo cross





# Configurazione disco e installazione DRBD

Partition table su ciascun nodo:

Device	Boot	Start	End	Blocks	Id	System
/dev/sda1		1	487	3911796	83	Linux
/dev/sda2		488	610	987997+	82	Linux swap
/dev/sda3		611	643	265072+	83	Linux
/dev/sda4		644	6693	48596625	83	Linux

# `mkfs`, etc...

NB: release 0.7 ha problemi con XFS per variable block size.

**sda3** è un metadisk di 128K per ogni partizione condivisa da drbd

**sda4** è la partizione usata da drbd per il mirror

Installazione di DRBD:

```
cd /usr/src
```

```
wget http://www.drbd.org/uploads/media/drbd-0.7_pre10_20040709.tar.gz
```

```
tar zxvf drbd-0.7_pre10_20040709.tar.gz
```

```
cd drbd-0.7_pre10_20040709
```

```
make
```

```
make install
```

# Configurazione drbd

```
cat > /etc/drbd.conf
```

```
resource drbd0 {
    protocol C;
    incon-degr-cmd "halt -f";
}

net {
    sndbuf-size    512k;
    timeout        60    ;
    connect-int    10    ;
    ping-int       10    ;
    ko-count       10    ;
}
```

```
on novanta.stat.unipd.it {
    device /dev/nb0;
    disk /dev/sda4;
    address 192.168.90.90:7789;
    meta-disk /dev/sda3[0];
}

on novantuno.stat.unipd.it {
    device /dev/nb0;
    disk /dev/sda4;
    address 192.168.90.91:7789;
    meta-disk /dev/sda3[0];
}

syncer {
    rate 512M;
}
}
```

# Far partire drbd

Lanciare drbd sulle due macchine  
`/etc/init.d/drbd start`

Sulla primaria (novanta):  
`drbdadmin /dev/nb0 primary`

Per controllare lo stato di avanzamento della  
sincronizzazione:  
`cat /proc/drbd`

# Installare e configurare heartbeat

```
apt-get install heartbeat  
ln -s /etc/init.d/drbd /etc/rc2.d/S19drbd
```

Startup automatico drbd

```
cat > /etc/ha.d/ha.cf  
logfacility local0  
keepalive 2  
deadtime 10  
initdead 120  
serial /dev/ttyS0  
baud 19200  
udpport 694  
udp eth1  
udp eth0  
node novantuno.stat.unipd.it  
node novanta.stat.unipd.it
```

tripla ridondanza: 2 ethernet  
e una seriale.

```
cat > /etc/ha.d/authkeys  
auth 3  
3 md5 PASSWORD
```

Nodi del cluster

# Script 'cluster' per init

```
cat > /etc/init.d/cluster
```

```
#!/bin/bash
# Date: Jul 2004
# Author: Alberto Cammozzo
# (mmzz - at - stat.unipd.it)
# Released under GNU GPL license
```

```
SHARED_MOUNTPOINT="/vservers"
```

```
DRBD="/etc/init.d/drbd"
```

```
DRBDADM="/sbin/drbdadm"
```

```
MOUNT="/bin/mount"
```

```
UMOUNT="/bin/umount"
```

```
VSERVER="/etc/init.d/vservers"
```

```
case "$1" in
    start)
        echo -n "Starting cluster:"
        $DRBDADM primary all && $MOUNT \
        $SHARED_MOUNTPOINT && $VSERVER start
        echo "."
        ;;
    stop)
        echo -n "Stopping cluster:"
        $VSERVER stop
        $UMOUNT $SHARED_MOUNTPOINT
        $DRBDADM secondary all
        $DRBDADM wait_connect all
        echo "."
        ;;
    *)
        echo "Usage: /etc/init.d/cluster
{start|stop}"
        exit 1
        ;;
esac

exit 0
```

# Vserver come servizio di heartbeat

Collocare i link nei posti necessari a heartbeat e a init (rc)

```
ln -s /etc/init.d/cluster /etc/ha.d/resource.d/cluster
```

Configurazione di heartbeat:

- **drbddisk** è uno script già installato da drbd
- **cluster** lo abbiamo scritto noi

```
cat > /etc/ha.d/haresources
novanta.stat.unipd.it drbddisk::drbd0 cluster
```

```
cat >> /etc/fstab
/dev/nb0 /vservers ext3 noauto 0 0
```

server primario

script per drbd

argomento

# Tuning vservers

```
cat >> /etc/vservers/pippo.sh
```

```
#!/bin/sh
# ---- see -- http://archives.linux-vserver.org/200406/0013.html
# ATTENZIONE: ARP diverso per server diverso
case $1 in
pre-start)
    /usr/lib/heartbeat/send_arp eth1 147.162.35.92 00:E0:18:02:C0:7D 147.162.35.255 ffffffff
    /usr/lib/heartbeat/send_arp eth1 147.162.35.92 00:E0:18:02:C0:7D 147.162.35.254 ffffffff
    ;;
post-start)
    ;;
pre-stop)
    ;;
post-stop)
    ;;
*)
    echo $0 pre-start
    echo $0 pre-stop
    echo $0 post-start
    echo $0 post-stop
    ;;
esac
```

**Viene eseguito prima di far partire il vserver**

**Gratuitous arp, necessario per aggiornare le tabelle dei router, client e altri server perchè non mandino pacchetti al vecchio server.**

# drbdadm

- drbdadm: si riferisce al file di configurazione
  - attach/detach: collega partizione a device
  - connect/disconnect: connette la rete
  - primary/secondary: attiva device in una modalità
  - invalidate/invalidate-remote: forza risincronizzazione
  - wait-connect: attende che il remoto si connetta
  - state: report
  - ...



# drbdsetup

- a basso livello. Non richiede file configurazione
  - net addr:port addr:port **protocol**
    - protocol (A,B,C): *write complete*
    - **A** : disco locale e buffer TCP locale
    - **B** : disco locale e buffer cache remota
    - **C** : disco locale e remoto
  - invalidate, invalidate\_remote
- drbdsetup /dev/nb0 disk /dev/sda4
- drbdsetup /dev/nb0 net 192.168.1.1 192.168.1.2 B
- --do-what-I-say

# DRBD

- Grazie a Lars Gunnar Ellenberg: l.g.e -at- web.de
- [www.drbd.org](http://www.drbd.org)
- **ATTENZIONE** è geloso (drbd, Lars non so):
  - non montare un secondary!
  - non accedere a un device sotto drbd senza di lui.
  - non funziona con LVM (LVM2 sì [?])
  - problemi con XFS

# heartbeat

- heartbeat su UDP + address takeover (*fake*)
- modello primary/secondary
- webservers/mail/ database/ firewall/ DNS/...
- [www.linux-ha.org](http://www.linux-ha.org)

# cosa manca?

- Analisi performance:
  - tuning richiede:  $MTA > \text{block size}$
- File system distribuiti
  - [openGFS.sourceforge.net](http://openGFS.sourceforge.net)
  - Intermezzo ([inter-mezzo.org](http://inter-mezzo.org))
  - [lustre.org](http://lustre.org)
- HA con bilanciamento dei carichi

Grazie