Introducción (I)

- Cuando se pretende garantizar cierto grado de redundancia de la información almacenada en un ordenador se utilizan sistemas en RAID:
  - RAID-1: Dos discos, siendo uno copia del otro.
  - RAID-5: Tres o más discos donde en un sector se almacena el XOR de los sectores correspondientes de los otros discos.
- Si deseamos que esta información se almacena de forma redundante en varios equipos podemos utilizar Distributed Replicated Block Device (DRBD).

Introducción (II)

- DRBD realiza RAID-1 entre los dispositivos de bloques de dos ordenadores (nodos) diferentes unidos por la red.
- Utiliza protocolo TCP/IP.
- Su funcionamiento se basa en que los nodos están en dos estados:
  - Primario: Puede acceder a los datos para leer y/o escribir.
  - Secundario: No puede acceder a los datos, solo copia los datos que le envía el otro nodo.

Configuración de DRBD (I)

- Se realiza en el fichero /etc/drbd.conf.
  - El fichero debe ser idéntico en ambos nodos.
- Actualmente solo contiene dos líneas:

```
include "drbd.d/global_common.conf";
include "drbd.d/*.res";
```

- Por tanto la información se almacena en ficheros dentro del directorio /etc/drbd.d.
- Los ficheros están formados por líneas que configuran secciones.
  - El carácter # indica el comienzo de un comentario y toda la línea a partir del mismo es ignorada.

Configuración de DRBD (II)

- La sintaxis de una sección es:

```
sección [nombre] { <parámetro> [valor]; [...] }
```

- Una sección puede contener en su interior otras secciones.
- Los parámetros comienzan con el nombre del parámetro seguido por espacios en blanco y los valores que toma el parámetro dentro de la sección, terminando con el carácter punto y coma.

Secciones de DRBD (I)

- DRBD posee las siguiente secciones:
- skip
  - Define una sección que es un comentario de más de una línea.
  - Todo el texto dentro de la sección skip es ignorado.
  - Se ignora también cualquier valor hasta el símbolo { de comienzo de sección.

```
skip resource drbd0 {
...
}
```

Secciones de DRBD (II)

- global
  - Configura parámetros globales.
  - Solo puede existir una sección global en el fichero de configuración.
  - Los parámetros permitidos en esta sección son:
    - minor-count
    - dialog-refresh
    - disable-ip-verification
    - usage-count

Secciones de DRBD (III)

- common
  - Define valores comunes para todos los recursos.
  - Estos valores pueden ser modificados por cada recurso de DRBD de forma particular.
  - Dentro de esta sección se permiten secciones:
    - startup
    - syncer
    - handlers
    - net
    - disk

Secciones de DRBD (IV)

- resource <nombre>
  - Configura un recurso de DRBD.
  - Es obligatorio que en su interior existan dos secciones on <nombre ordenador>.
  - Puede tener en su interior opcionalmente otras secciones:
    - startup
    - syncer
    - handlers
    - net
    - disk

Secciones de DRBD (V)

- on <nombre ordenador>
  - Configura uno de los nodos del dispositivo DRBD.
  - <nombre ordenador> debe coincidir con el nombre obtenido mediante el comando uname -n.
  - Requiere la especificación de los parámetros:
    - device
    - disk
    - address
    - meta-disk
- disk
  - Especifica el comportamiento del sistema respecto al dispositivo de bloques del disco.
  - Algunos parámetros son on-io-error y resync-rate.

Secciones de DRBD (VI)

- net
  - Define opciones de configuración de la red.
  - Sus parámetros opcionales son:
    - sndbuf-size
    - rcvbuf-size
    - timeout
    - connect-int
    - ping-int
    - ping-timeout
    - max-buffers
    - max-epoch-size
    - ko-count

Secciones de DRBD (VII)

- startup
  - Define el funcionamiento en el arranque.
  - Sus parámetros opcionales son:
    - wfc-timeout
    - degr-wfc-timeout
    - outdated-wfc-timeout
- handlers
  - Permite definir ejecutables que serán arrancados por DRBD en respuesta a ciertos eventos.

Parámetros de DRBD (I)

- minor-count <valor>
  - Número de dispositivos que pueden ser definidos una vez rearrancado el servicio.
- dialog-refresh <tiempo>
  - Tiempo de refresco del dialogo con el usuario.
- disable-ip-verification
  - Indica que DRBD no verifique la dirección de red del sistema.

Parámetros de DRBD (II)

- protocol <identificador>
  - Tipo de protocolo utilizado en la comunicación.
    - A: Se consideran escritos los datos al disco si se han escrito en el disco local y enviados al buffer TCP/IP.
    - B: Se consideran escritos los datos al disco si se han escrito en el disco local y recibidos por el nodo remoto.
    - C: Se consideran escritos los datos al disco si se han escrito en disco por ambos nodos.
  - El orden de seguridad es C > B > A.
  - El orden de rapidez es A > B > C.

Parámetros de DRBD (III)

- device <nombre> minor <número>
  - Nombre del dispositivo de bloques de DRBD.
  - Es posible omitir <nombre> y se utiliza /dev/drbd<número>
  - Es posible omitir minor <número> y se utiliza el indicado por <nombre>.
- disk <nombre>
  - Es el nombre del dispositivo del nodo (dev/sda1 por ejemplo).
- address <dirección IP:puerto>
  - Dirección IP y puerto a utilizar por DRBD en el nodo.

Parámetros de DRBD (IV)

- meta-disk {internal, dispositivo[indice]}
  - DRBD utiliza metadatos para decidir que debe sincronizar, etc.
  - Los metadatos utilizan 128 MBytes.
  - internal: Los metadatos se almacenan al final del propio dispositivo del nodo especificado por disk.
  - dispositivo[indice]: Los metadatos se almacenan en el dispositivo indicado.
    - Indice indica el desplazamiento, en unidades de 128 MB respecto al origen del dispositivo.
    - Permite almacenar en un dispositivo metadatos de varios dipositivos.

Parámetros de DRBD (V)

- on-io-error {pass_on, call-local-io-error, detach}
  - Acción a realizar en caso de error de lectura o escritura.
  - pass_on: Informar del error al dispositivo DRBD.
  - call-local-io-error: Ejecutar el comando indicado por local-io-error.
  - detach: Continuar el funcionamiento normalmente.

Parámetros de DRBD (VI)

- sndbuf-size <tamaño>
  - Tamaño del buffer de envío del socket TCP.
  - Valor por defecto de 128 Kbytes.
- rcvbuf-size <tamaño>
  - Tamaño del buffer de recepción del socket TCP.
  - Valor por defecto de 128 Kbytes.

Parámetros de DRBD (VII)

- timeout <tiempo (en décimas)>
  - Tiempo de espera de la respuesta a un paquete antes de considerar al otro nodo no activo y cerrar la conexión TCP/IP
  - El valor por defecto es de 60 (6 segundos).
  - Debe ser menor que connect-int y ping-int.
- connect-int <tiempo (en segundos)>
  - Tiempo que transcurren entre dos intentos de conexión.
  - El valor por defecto es 10 segundos.

Parámetros de DRBD (VIII)

- ping-int <tiempo (en segundos)>
  - Tiempo de espera antes de enviar un paquete para confirmar que el otro nodo sigue activo.
  - El valor por defecto es de 10 segundos.
- ping-timeout <tiempo (en milésimas de segundo)>
  - Especifica el tiempo que espera la respuesta de un paquete enviado por ping-int antes de considerar al otro nodo muerto.
  - El tiempo por defecto es de 500 milisegundos.

Parámetros de DRBD (IX)

- max-buffers <número>
  - Número de buffers reservados por DRBD.
  - El valor por defecto es 32 buffers de 4 Kbytes.
- ko-count <valor>
  - Número de veces que debe fallar el nodo secundario en una escritura para ser excluido del cluster.
  - El valor por defecto 0 deshabilita esta opción.
- max-epoch-size <número>
  - Máximo número de bloques de datos entre dos escrituras.

Parámetros de DRBD (X)

- wfc-timeout <tiempo (segundos)>
  - Bloquea el arranque del nodo los segundos indicados esperando el arranque del otro nodo.
  - El valor 0, valor por defecto, indica una espera infinita.
- degr-wfc-timeout <tiempo (segundos)>
  - Bloquea el arranque del nodo los segundos indicados esperando el arranque del otro nodo si el sistema estaba degradado al apagarlo.
- outdated-wfc-timeout <tiempo (segundos)>
  - Tiempo de espera en el arranque si el otro nodo esta desactualizado.

Parámetros de DRBD (XI)

- resync-rate <valor> (rate <valor>)
  - Ancho de banda utilizado por DRBD en la sincronización de datos.
  - El valor por defecto es de 250 KBytes/s.
  - El valor máximo es de 700000K
- local-io-error <comando>
  - Comando a ejecutar si se produce un error en la entrada/salida del subsistema local.

Ejemplo de configuración (I)

- Determinar donde almacenar los metadatos:
  - internal: Redimensionar el tamaño de la partición.

```
resize2fs <partición> <tamaño final>
```

  - dispositivo[indice]: Utilizaremos esa partición.

Ejemplo de configuración (II)

```
resource drbd0 {
   net { protocol C; }
   syncer { resync-rate 10M;}
   startup { wfc-timeout 300;
             degr-wfc-timeout 150;
             iutdated-wfc-timeout 150;}
   on nodo1 { device /dev/drbd0;
              disk /dev/sda2;
              address 192.168.0.1:7780;
              meta-disk internal;}
   on nodo2 { device /dev/drbd0;
              disk /dev/sda2;
              address 192.168.0.2:7780;
              meta-disk internal; }
}
```

Ejemplo de configuración (III)

- El área de metadatos debe inicializarse utilizando, en cada nodo, el comando:

```
drbdadm create-md <nombre del recurso>
```

Arranque y comprobación (I)

- El servicio de DRBD se arranca como:

```
service drbd start
```

- Su estado se comprueba como:

```
service drbd status
m:res    cs         st                 ds                 p   mounted      fstype
0:drbd0  Connected  Secondary/Secondary  UpToDate/UpToDate  C
```

- Para montar el sistema de ficheros en un nodo debe ponerse en primario:

```
drbdadm primary drbd0
```

- Y montarlo:

```
mount /dev/drbd0 /directorio
```

Arranque y comprobación (II)

- El estado será en el nodo que lo tiene montado:

```
m:res    cs         st                 ds                 p   mounted       fstype
0:drbd0  Connected  Primary/Secundary  UpToDate/UpToDate  C   /directorio   ext3
```

- Y en el otro nodo:

```
m:res    cs         st                 ds                 p   mounted       fstype
0:drbd0  Connected  Secondary/Primary  UpToDate/UpToDate  C
```

Sincronización de los nodos (I)

- Inicialmente los nodos están desincronizados

```
m:res    cs         st                    ds                          p   mounted      fstype
0:drbd0  Connected  Secondary/Secondary   Inconsistent/Inconsistent   C
```

- Forzar la sincronización inicial usando, en el nodo que contenga la información correcta:

```
drbdadm primary --force drbd0
```

Sincronización de los nodos (II)

- En ciertos casos de error pueden desincronizarse.

```
m:res    cs         st                 ds                 p   mounted       fstype
0:drbd0  StandAlone Secondary/Unknown  UpToDate/DUnknow   C
```

- Forzar la sincronización indicando que nodo es erróneo y luego forzar la conexión.

```
drbdadm invalidate drbd0
drbdadm connect drbd0
```

- Obteniendo:

```
m:res    cs         st                 ds                   p   mounted       fstype
0:drbd0  SyncSource Secondary/Secondary  UpToDate/Inconsistent  C

m:res    cs         st                 ds                   p   mounted       fstype
0:drbd0  SyncTarget Secondary/Secondary  Inconsistent/UpToDate  C
```