LinuxHA+DRBD: Alta disponibilidad de bajo coste

Jesús Espino García
jespinog@gmail.com

¿Qué es clustering?

- Conjunto de maquinas.
- Trabajando en equipo.
- Con un mismo objetivo.
- Los casos mas habituales son:
  - Cluster computacional.
  - Alta Disponibilidad.
  - Balanceo de carga.

Alta Disponibilidad VS Toleraciona a Fallos

- Tolerancia a fallos:
  - Caro.
  - Hardware tolerante a fallos (redundancia interna, detección de errores).
  - Perdidas de servicio muy breves, entre décimas de segundo y unos pocos segundos.
- Alta disponibilidad:
  - Barato.
  - Redundancia de hardware.
  - Perdida de servicios entre algunos segundos hasta unos pocos minutos.
  - Hace uso en subsistemas tolerantes a fallos.

¿Qué es Linux HA?

- Un sistema de clustering de Alta disponibilidad.
- Libre.
- Maduro: En el mercado desde 1998.
- Extendido: Mas de 30.000 cluster en producción.
- Una comunidad de desarrollo abierta liderada por IBM y Novell.
- Distribuido con la mayoría de las distribuciones.
- Sin requerimientos de hardware especial.
- Independiente del kernel (todo en espacio de usuario).

¿Qué nos permite hacer?

- Elaborar sistemas resistentes a un único punto de fallo.
- Los tiempos de caída son muy cortos.
- Aproximadamente añade un "9" a tu disponibilidad.
- Hasta 16 nodos.
- Recuperación en caso de fallo de nodo o de servicio.
- Recuperación en caso de perdida de conectividad.
- Clusters activo-pasivo y activo-activo.
- Monitorización de recursos.
- Interfaz gráfico de configuración y monitorización.

¿Qué no puede hacer?

- Dar el 100 % de disponibilidad (nadie puede).

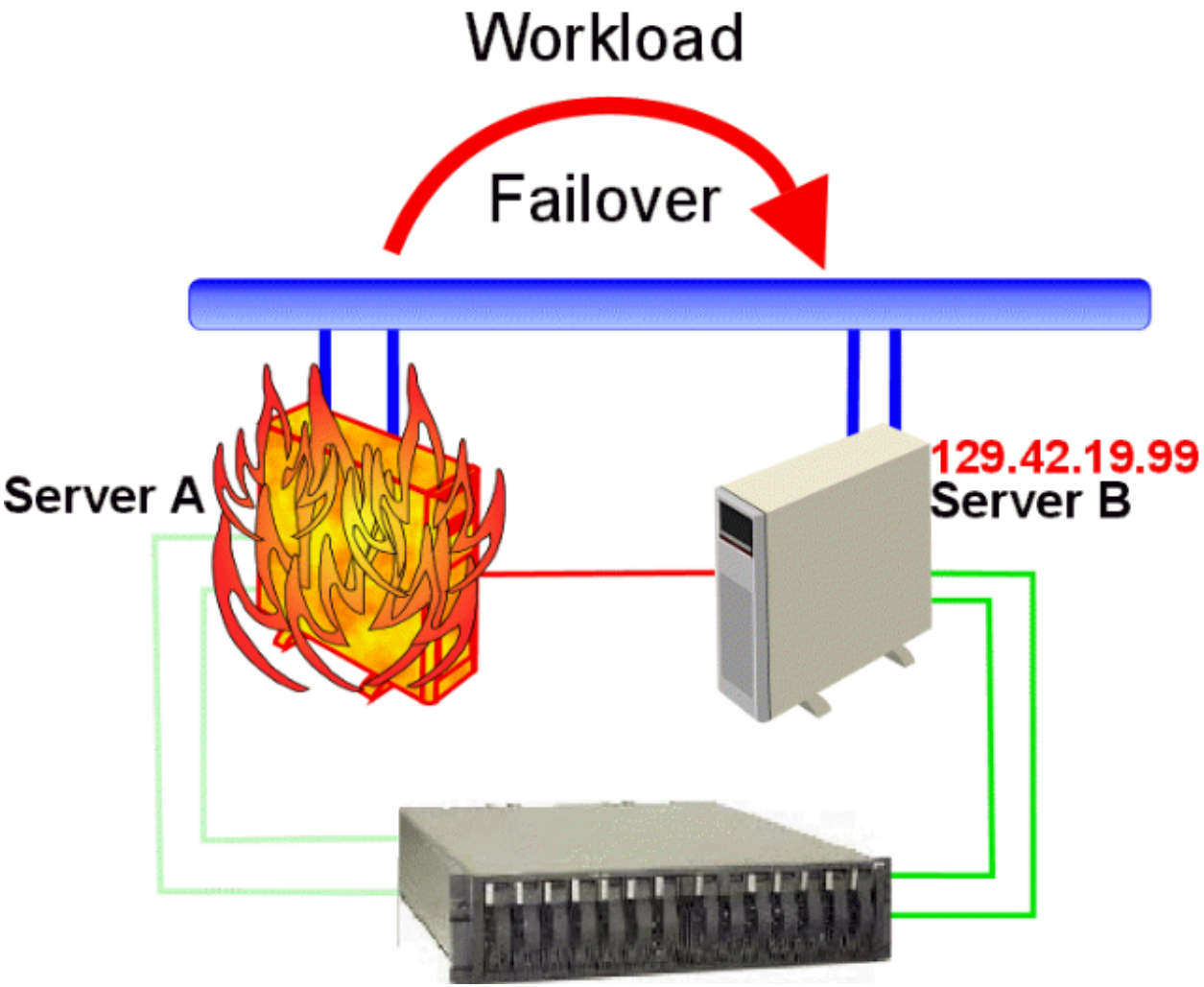
Algunos numeros

- 99,9999 % → 30sec
- 99,999 % → 5min
- 99,99 % → 52min
- 99,9 % → 9h
- 99 % → 3.5días

Algunos conceptos

- Nodo: Maquina (real o virtual) que forma parte del cluster.
- Recurso: Algo que manejar (Servicio, IP, Unidad de disco...).
- Agente de recurso: Un (init) script que controla un recurso.
- Heartbeat: Pulso de comunicación que indica a un nodo el estado de los demás.
- Failover: Toma de control por parte de un nodo de los servicios de otro.
- Split Brain: Falla la comunicación "Heartbeat" entre dos grupos diferentes de nodos.
- Quorum: Sistema de acuerdos en el cluster.
- Fencing: Exclusion de un nodo del cluster.
- SPOFs (Single Point of Failure).

Failover



Split Brain

- División del cluster por culpa de fallos.
- Normalmente de comunicación del Heartbeat.
- Cada nodo crea ser el nodo activo.
- OJO!! Peligro de corrupción de datos.

Quorum

- Es un intento de evitar la mayor parte de los fallos de split brain.
- Trata de asegurase que solo uno de las divisiones este activa.
- Uno de los sistemas mas comunes de quorum es el voto. Solo la división con mas nodos puede ejecutar el cluster.
  - Esto no funciona muy bien para 2 nodos.
- Otro sistema de quorum es el acceso a un recurso exclusivo, el que llegue después, desiste y cede los servicios.

Fencing

- Excluye a los nodos que estén dando fallos para que no accedan a los recursos del cluster.
- Linux HA soporta:
  - STONITH (Shoot The Other Node In The Head) para node Fencing.
  - Sistemas Self-Fencing (Sistemas que aseguran acceso exclusivo a un recurso (ServerRAID de IBM)) para resource fencing.

SPOFs

- Elemento del sistema que en caso de fallo puede dejar todo el sistema inutilizado.
- Un buen diseño elimina todos los SPOFs.
- Existe un limite, normalmente impuesto por el coste, en la eliminación de SPOFs.
- Mucho cuidado con los SPOFs no evidentes:
  - Unidades raid con una sola controladora.
  - Enlaces redundantes usando el mismo cable.

Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:

Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:
  - Redundancia.

Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:
  - Redundancia.
  - Redundancia.

Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:
  - Redundancia.
  - Redundancia.
  - Redundancia.

Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:
  - Redundancia.
  - Redundancia.
  - Redundancia.
- La mayoría de los SPOFs se eliminan con redundancia.

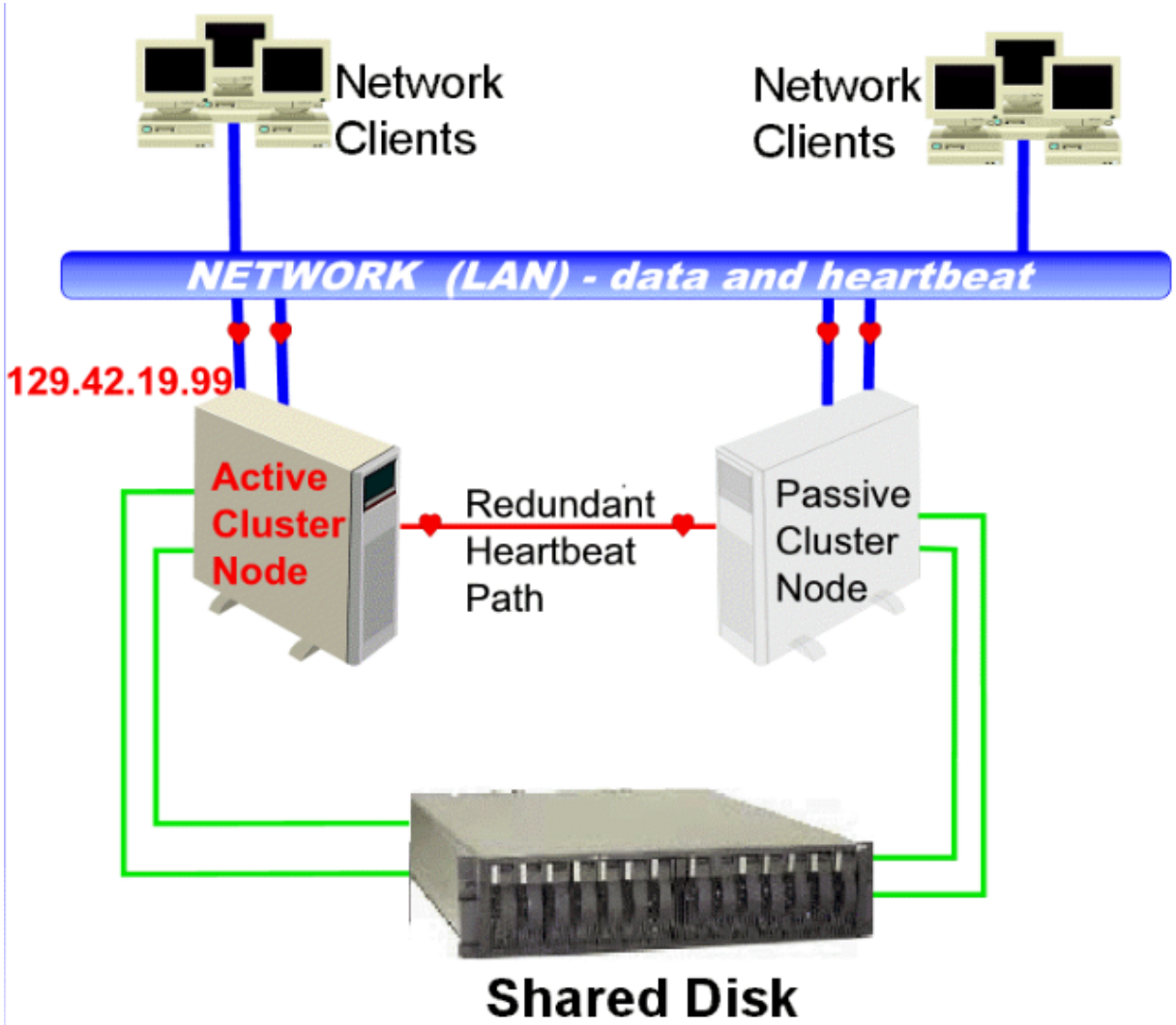
Sistemas NON-SPOFs

- Importante las tres R de la alta disponibilidad:
  - Redundancia.
  - Redundancia.
  - Redundancia.
- La mayoría de los SPOFs se eliminan con redundancia.
- Siempre existen algún punto de fallo simple, hay que saber donde parar.

Compartiendo datos: Mediante unidad compartida

- El modo normal.
- Eleva el precio de cluster.
- Da muy buen rendimiento.
- Solución de bajo coste (AoE).

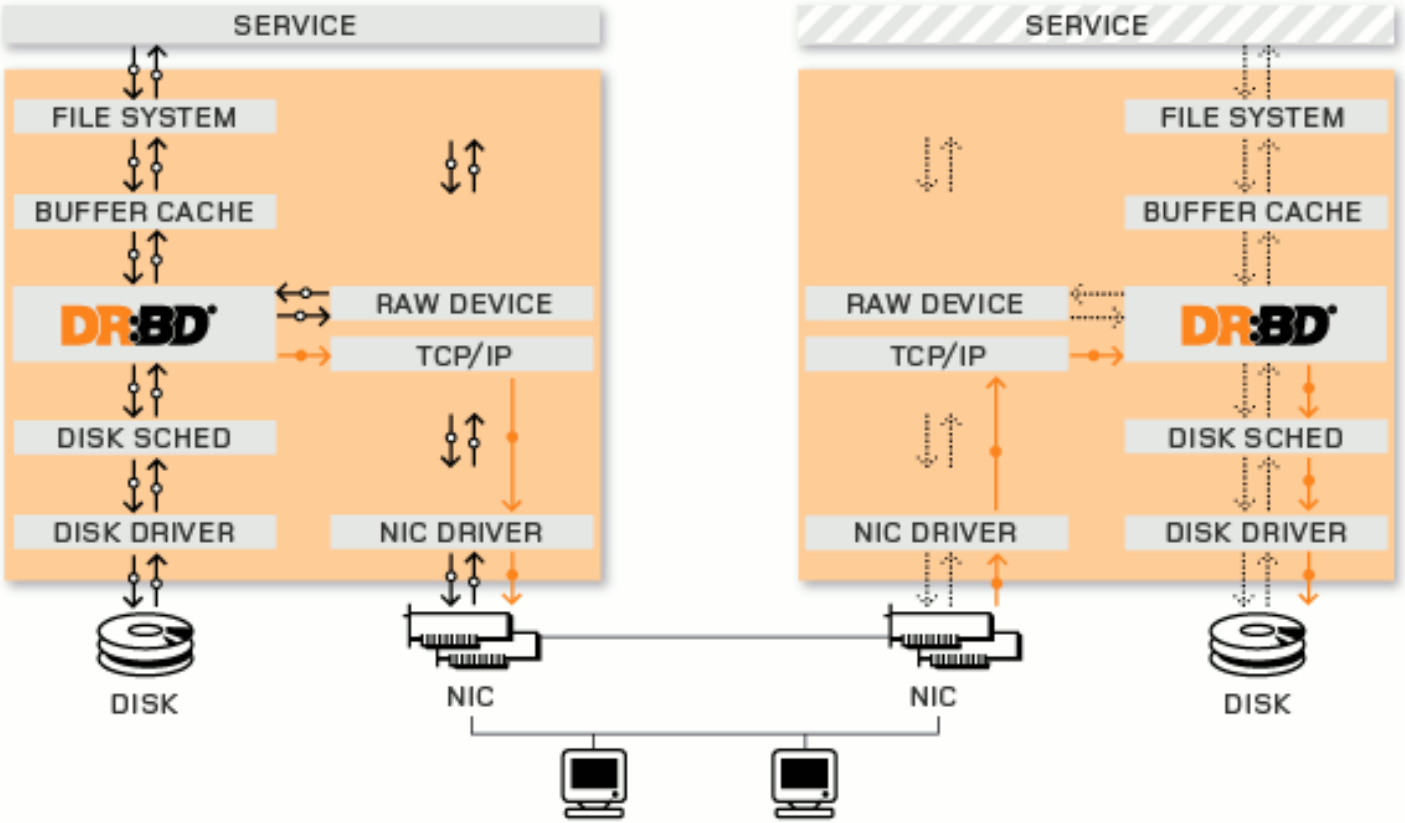
Compartiendo datos: Mediante unidad compartida



Compartiendo datos: Replicación de los datos

- El modo barato.
- No necesitas unidad compartida.
- El rendimiento es menor.
- Se basa en los sistemas de replicación de los propios servicios y en servicios de replicación de ficheros (DRBD).
- Una relación calidad/precio excelente.

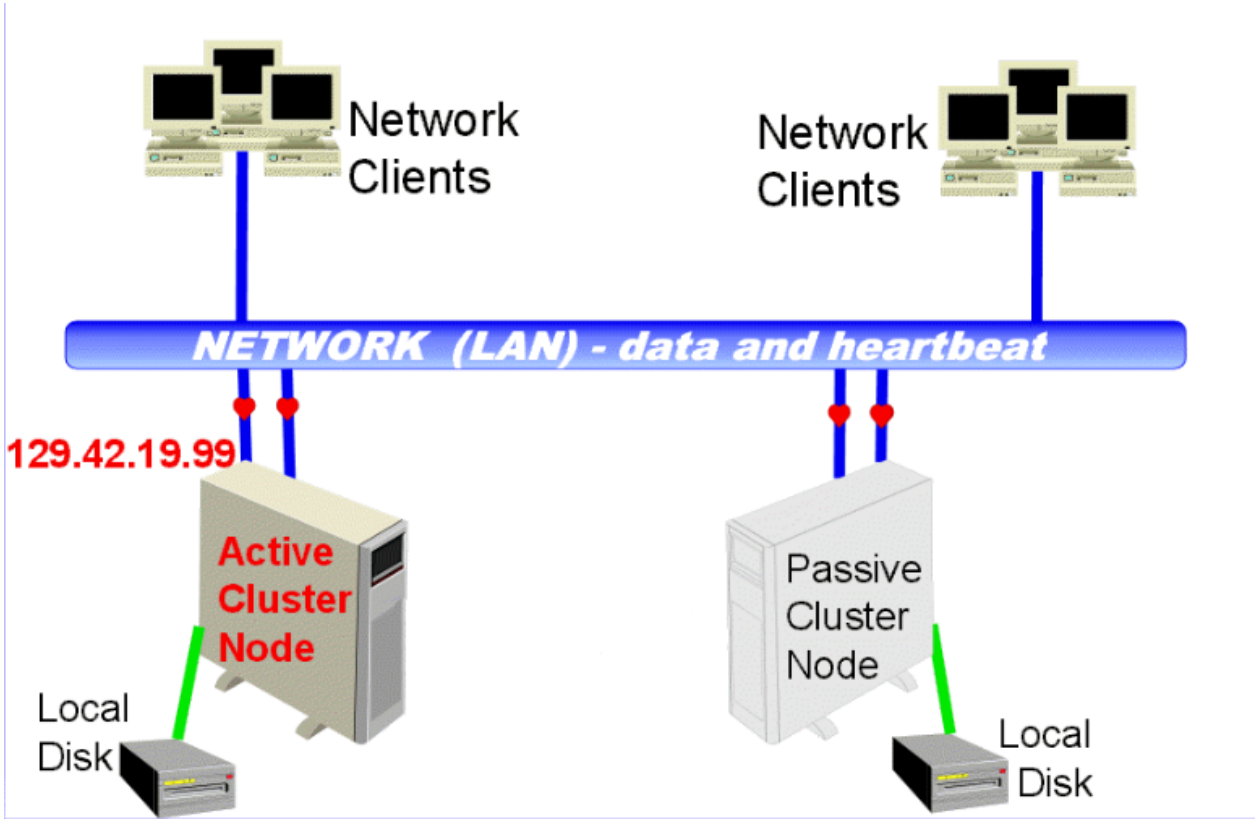
Compartiendo datos: Replicación de los datos



Compartiendo datos: Sin datos que compartir

- Es raro, pero existen casos.
- Firewalls.
- Balanceadores de carga.
- Servidores Proxy.
- Servidores Web estáticos (replicación inicial de los contenidos).

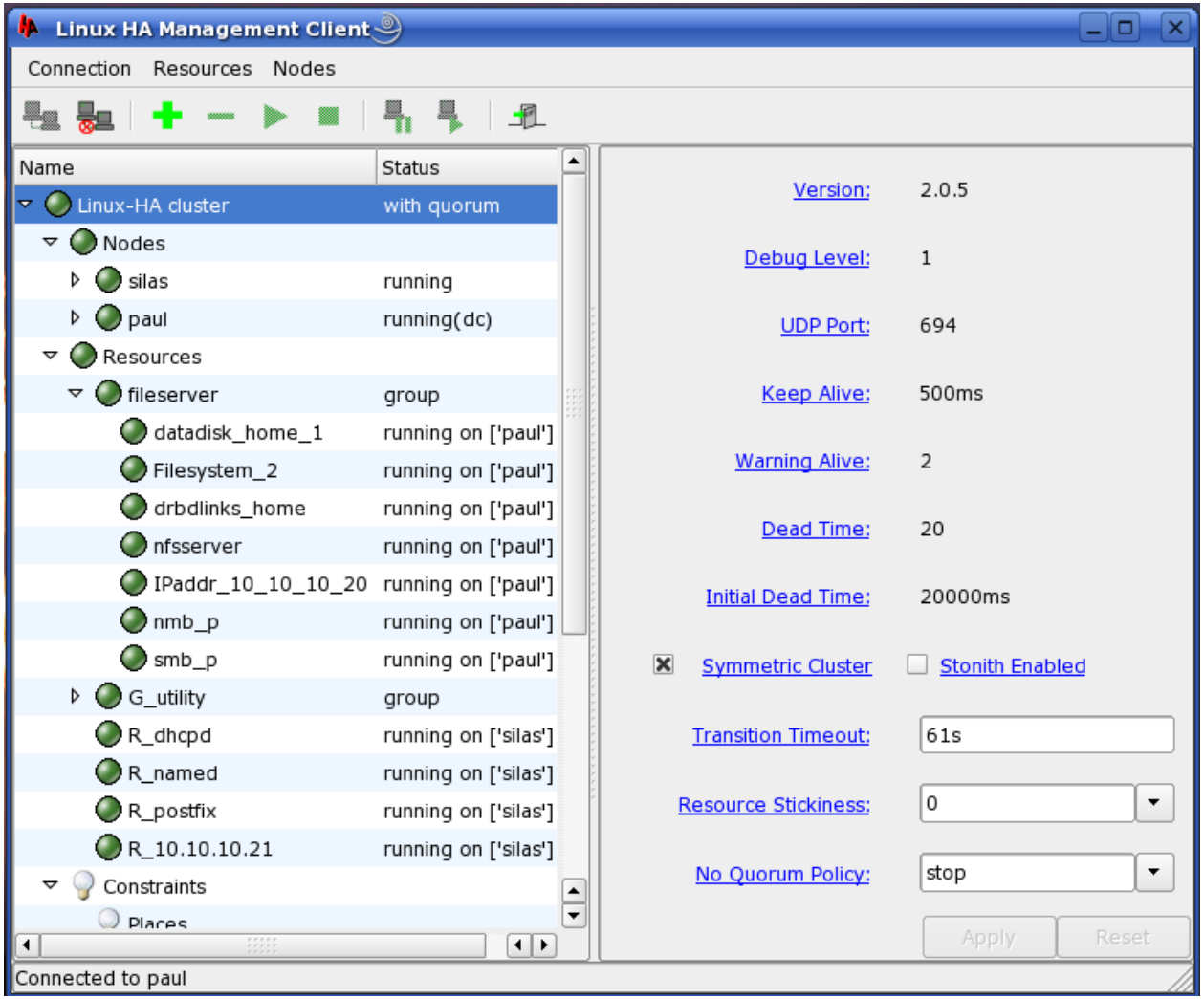
Compartiendo datos: Sin datos que compartir



Sistemas de ficheros de Cluster

- Preparados para accesos concurrentes desde varios nodos.
- Pensados para mantener la consistencia de los datos.
- Los sistemas de ficheros para cluster libres mas conocidos son:
  - GFS: Global File System (Comprado por RedHat y licenciado como GPL).
  - OCFS2: Oracle Cluster File System (Desarrollado por Oracle, GPL).

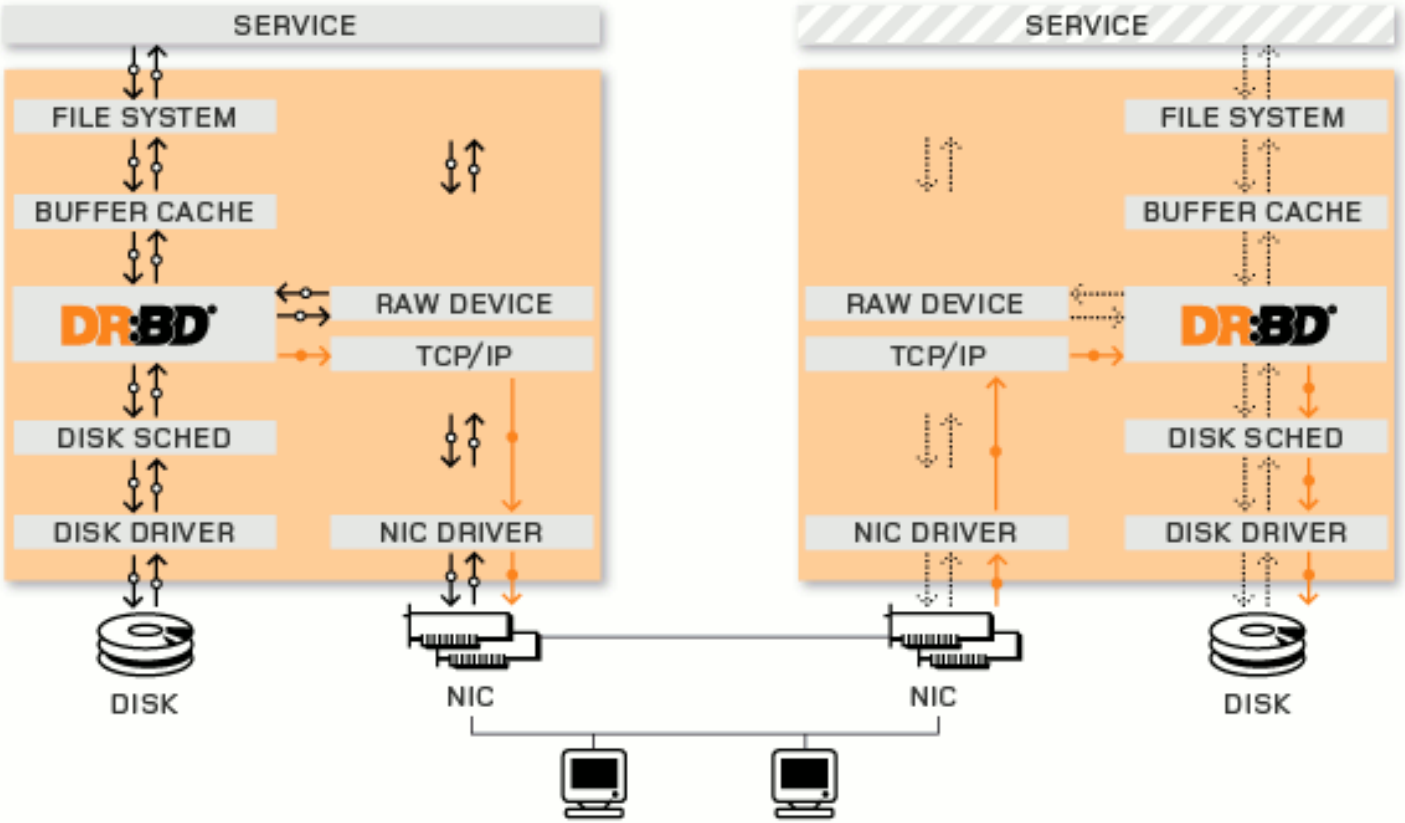
Interfaz gráfica



¿Qué es DRBD?

- Servicio de replica de dispositivos.
- GPL.
- Funcionamiento sobre dispositivos fisicos o virtuales.
- Funcionamiento bajo dispositivos virtuales (LVM).
- Sincronizacion Activo-Pasivo y Activo-Activo.
- Modos Asincrono, Semi-Sincrono y Sincrono.

¿Como funciona DRBD?



Ejemplo en vivo

- Montar dispositivo con sincronizacion DRBD.
- Montar PeaceMaker.
- Montar servicios sobre DRBD y PeaceMaker.

Referencias

- Linux-HA: http://www.linux-ha.org
- DRBD: http://www.drbd.org
- Comunidad: http://linux-ha.org/ContactUs (Listas, IRC, bugtracking...)
- GFS: http://sources.redhat.com/cluster/gfs/
- OCFS2: http://oss.oracle.com/projects/ocfs/

Dudas y preguntas

Dudas. . .