

DRBD 9 + Pacemaker HA

Almacenamiento de alto rendimiento y alta disponibilidad con software libre



XTEC Xarxa Telemàtica
Educativa de Catalunya

*Joan Estrada Campañá, IES Escola del Treball de Barcelona
Alberto Larraz Dalmases, XTEC (Xarxa Telemàtica Educativa de Catalunya)*



Índice

- Introducción
 - Necesidades de almacenamiento
- Almacenamiento y Caché
- Replicación y Redundancia
- Alta Disponibilidad
- Resultados



Escola del Treball, Barcelona

- 50 titulaciones
- 3500 usuarios
- 1200 máquinas
- Software Libre
- Infraestructura DiY





Isard VDI

Virtual Desktops Infrastructure

+ escritorios
+ almacenamiento

+ datos...
¿Cómo crecemos?

Hardware...

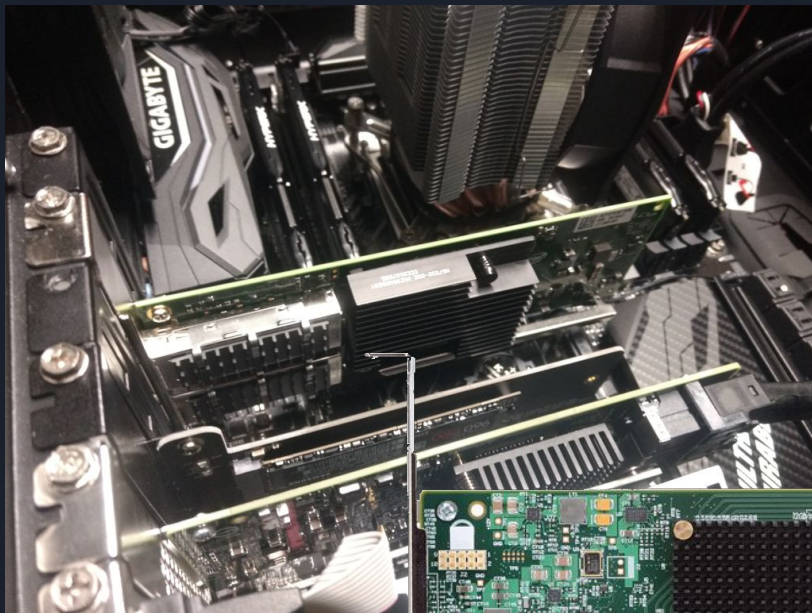


Intel 40Gbps
Dell Switch 40Gbps



NVME Samsung EVO y Pro

Hardware...





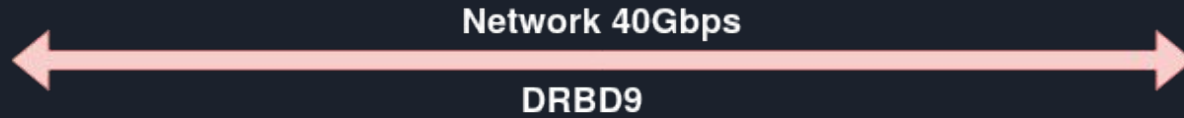
Stack de tecnologies



dm-writeboost



Almacenamiento y Caché: Diagrama



WRITE CACHE



RAID



WRITE CACHE



RAID

...



WRITE CACHE



RAID



Implementación

DRBD 9

- v8 (3 nodos) → v9 hasta (16 nodos)
- Replicación de dispositivos de bloques por red (kernel)
 - Según LINBIT (x6 CEPH *)
- Nodos síncronos, asíncronos, “clientes diskless” y de control



* Fuente: LINBIT
<https://www.linbit.com/en/ceph/>

Replicación y Redundancia: Configuración

```
vgcreate drbdpool <pvs>  
drbdmanage init <ip>  
drbdmanage add-node <name> <ip>  
drbdmanage add-resource <name>  
drbdmanage add-volume <resource> <size>  
drbdmanage deploy <resource> <number of nodes>
```

<http://techdocs.escoladeltreball.org/storage/drbd/>

Almacenamiento y Caché



Write cache (dm-writeboost)

NVME: Recibe todo el impacto de escritura

HD: Lentamente se vuelcan los datos

<http://techdocs.escoladeltreball.org/storage/cache/writeboost/>



Alta Disponibilidad

Pacemaker (HA en Linux)

- Múltiples recursos
 - Puntos de montaje
 - Servidor NFS
 - IP flotante
- Creamos recursos propios:
 - Para write caché

<http://techdocs.escoladeltreball.org/clusters/pacemaker/>

Alta Disponibilidad: Diagrama



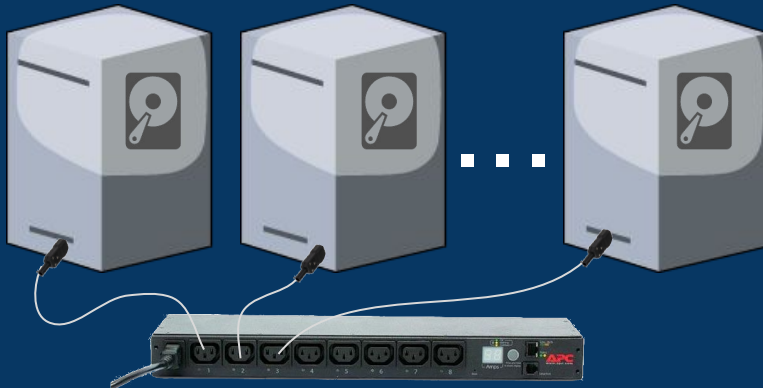
- Clientes montan almacenamiento desde ip flotante.

10.0.0.10 IPs flotantes

drbd node1

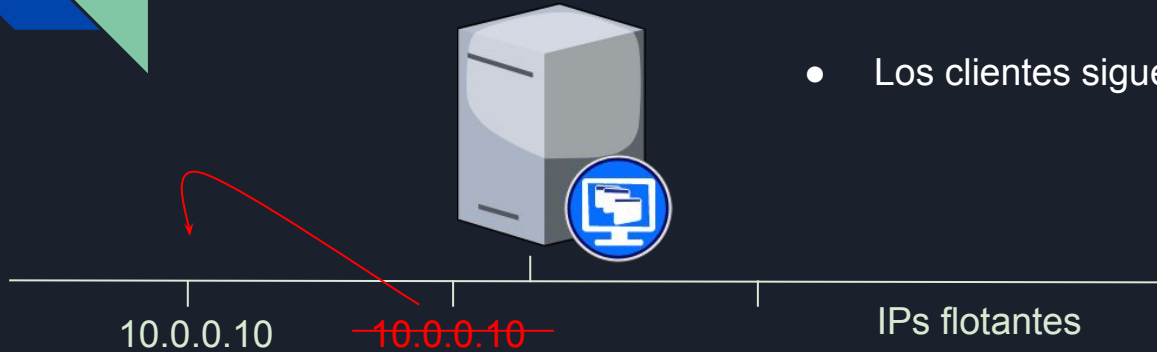
drbd node2

drbd node3

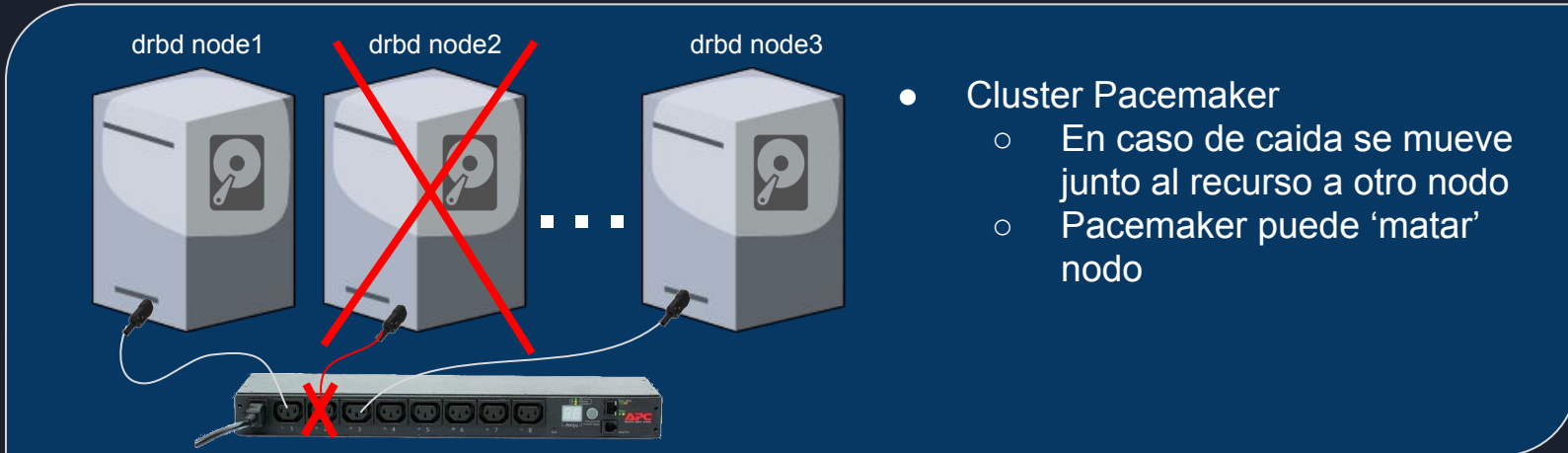


- Cluster Pacemaker
 - Recurso IP flotante
 - STONITH: Nodos conectados a red eléctrica por regleta IP

Alta Disponibilidad: Diagrama

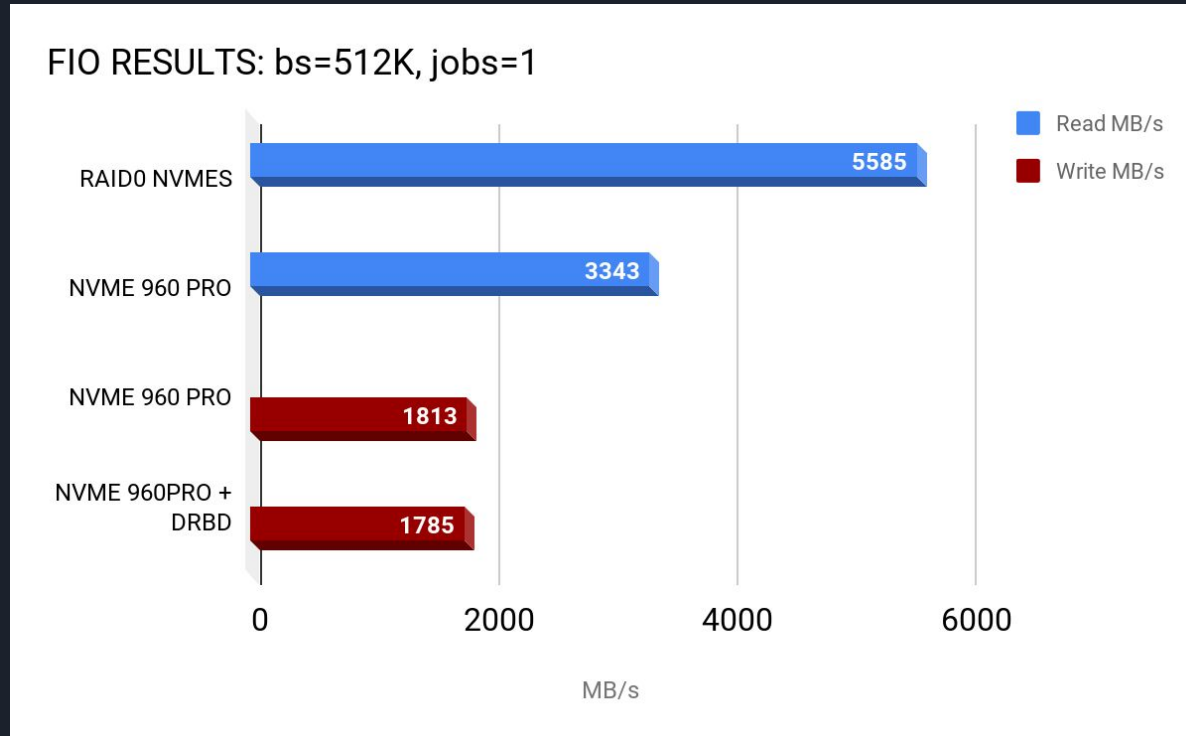


- Los clientes siguen trabajando



- Cluster Pacemaker
 - En caso de caída se mueve junto al recurso a otro nodo
 - Pacemaker puede 'matar' nodo

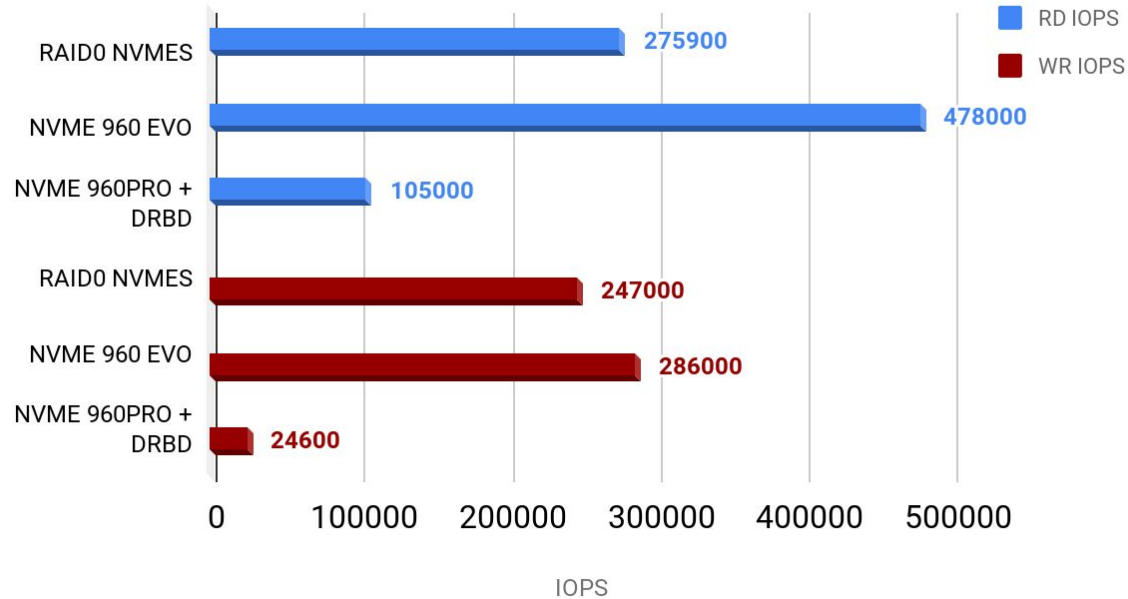
Resultados FIOS



<http://techdocs.escoladeltreball.org/utilities/fio/>

Resultados FIOS

FIO RESULTS: bs=4K, jobs=24



<http://techdocs.escoladeltreball.org/utilities/fio/>

Conclusiones

Evolución tecnológica

IOPS disco
ANCHO BANDA disco
RED



Protocolo TCP/IP



Network Performance

Protocolo TCP: genera latencias importantes

Velocidad de CPU: influye en el rendimiento

Hay que “tunear” parámetros y hacer FIOS



Subir GHz: CPU governor to 'performance'

```
cpufreq-set -r -g performance
```

TCP buffer: tamaños buffers y

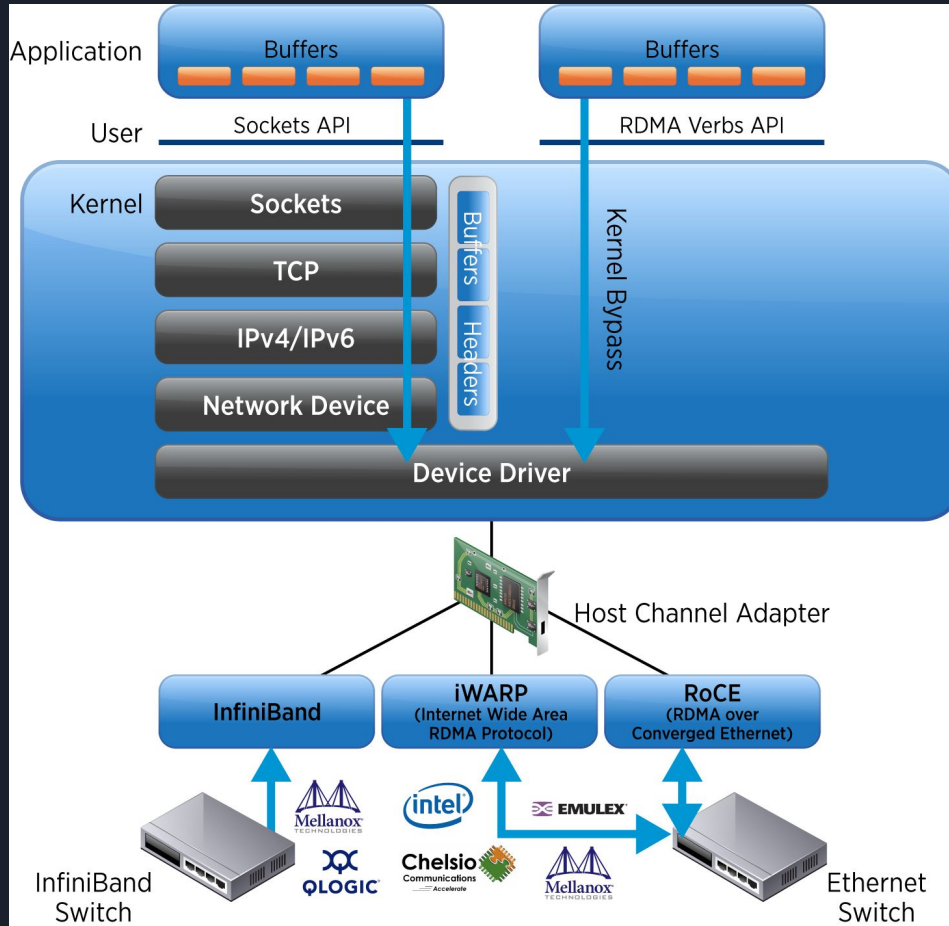
```
net.ipv4.tcp_rmem = 4096 87380 67108864
```

Mapeo estático o dinámico de IRQs para balancear cargas entre las CPU disponibles

Activar **'fair queuing' (FQ)** en las tarjetas de red

```
net.core.default_qdisc = fq
```

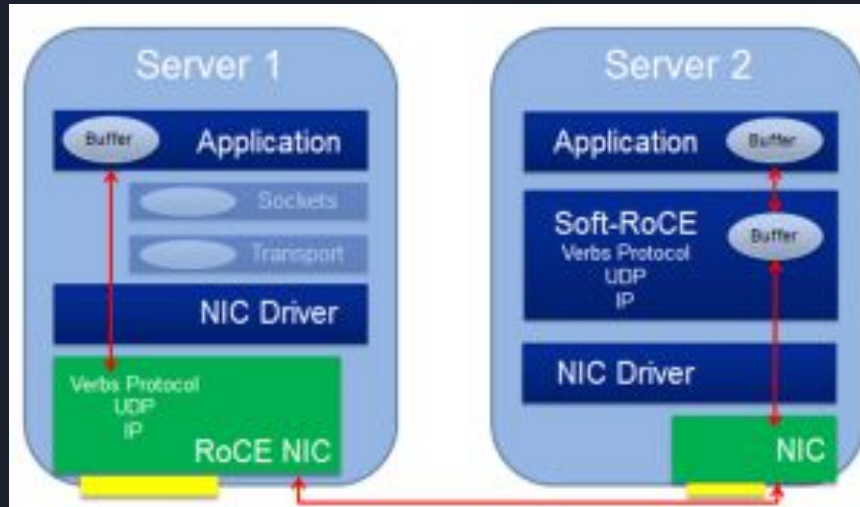
RDMA (remote direct memory access)



NVME 960 EVO		286000
NVME 960PRO + DRBD		24600

RoCE: RDMA over ethernet:

- Hardware RoCE
- Soft RoCE (modprobe rdma_rxe)



Documentación en techdocs.escoladeltreball.org



DRBD



Search



isard-vdi/thedocs
1 Stars · 0 Forks

IsardVDI Technical docs

Introduction

Storage ^

Concepts

Tools

Raid

DRBD

Cache v

Networking v

Clusters & HA v

Virtualization v

Setups v

Utilities v

About

Configure drbd9 cluster

Create drbdpool volume group on desired PVs

```
vgcreate drbdpool /dev/nvme0n1 /dev/mapper/disks
```

Initialize the drbdmanage in the master node with own parameters

```
drbdmanage init <node-name> <ip_drbd>
```

Add volume and resources

```
drbdmanage add-volume <volume-name> <capacity>
```

Note: It will create and associated resource with the same volume-name.

Deploy resources to nodes

```
drbdmanage deploy <resource-name> <nº nodes>
```

Just if you want to allocate volumes on desired PV

Table of contents

Install required packages

Configure drbd9 cluster

Types of nodes

Control node

Pure controller node

Satellite node

Pure client node

Utils



BRAINUPDATERS



INSTITUT
ESCOLA DEL TREBALL
DE BARCELONA

XTEC Xarxa Tècnica d'Educació de Catalunya

Gracias por la atención

Guías de instalación, configuraciones, tests...

<http://techdocs.escoladeltreball.org>

<https://github.com/isard-vdi>

sysadmin@escoladeltreball.org

