DRBD 9 + Pacemaker HA

Almacenamiento de alto rendimiento y alta disponibilidad con software libre

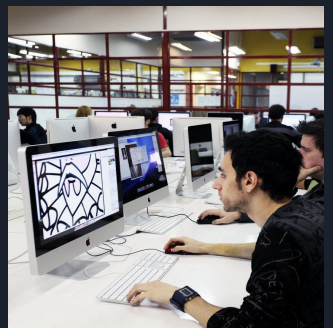
***Joan Estrada Campañá**, IES Escola del Treball de Barcelona*
***Alberto Larraz Dalmases**, XTEC (Xarxa Telemàtica Educativa de Catalunya)*

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?

Institut Escola del Treball de Barcelona

.XTEC Xarxa Telemàtica Educativa de Catalunya

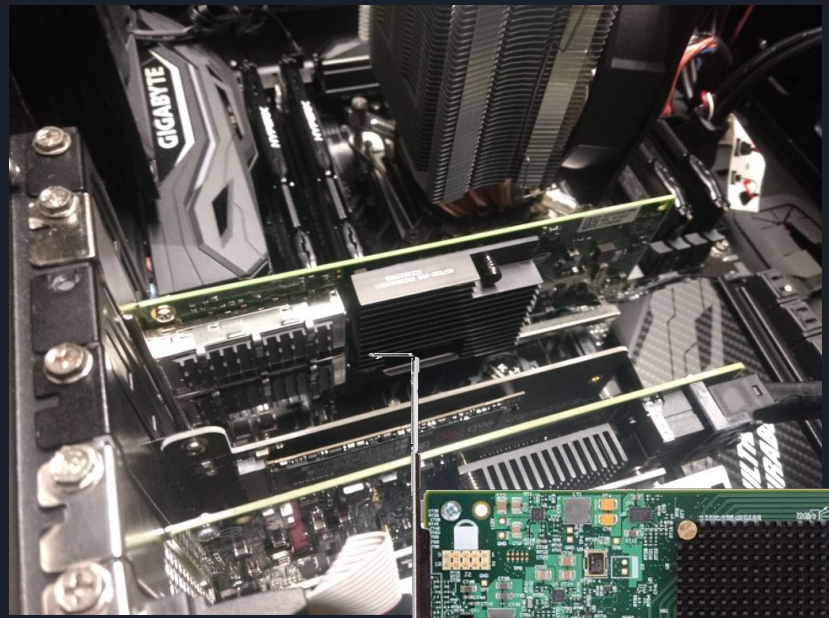
Hardware...

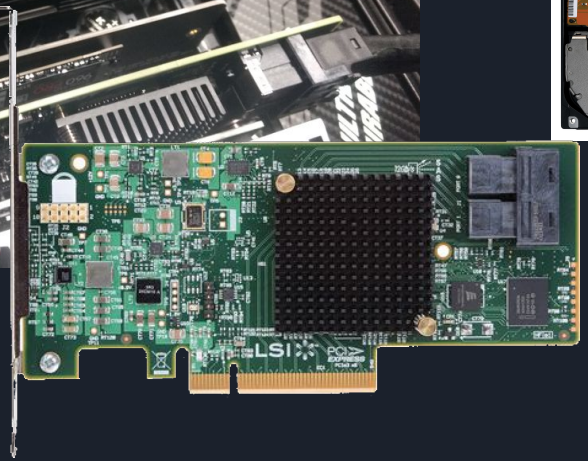
Intel 40Gbps

Dell Switch 40Gbps

NVME Samsung EVO y Pro

Hardware...





Stack de tecnologías

dm-writeboost

Almacenamiento y Caché: Diagrama

Network 40Gbps

DRBD9

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

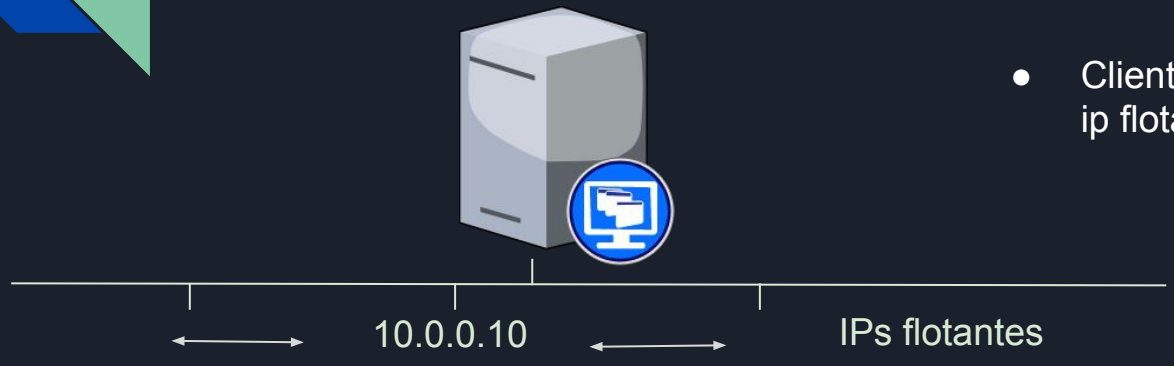
Pacemaker (HA en Linux)

- Múltiples recursos
  - Puntos de montaje
  - Servidor NFS
  - IP flotante

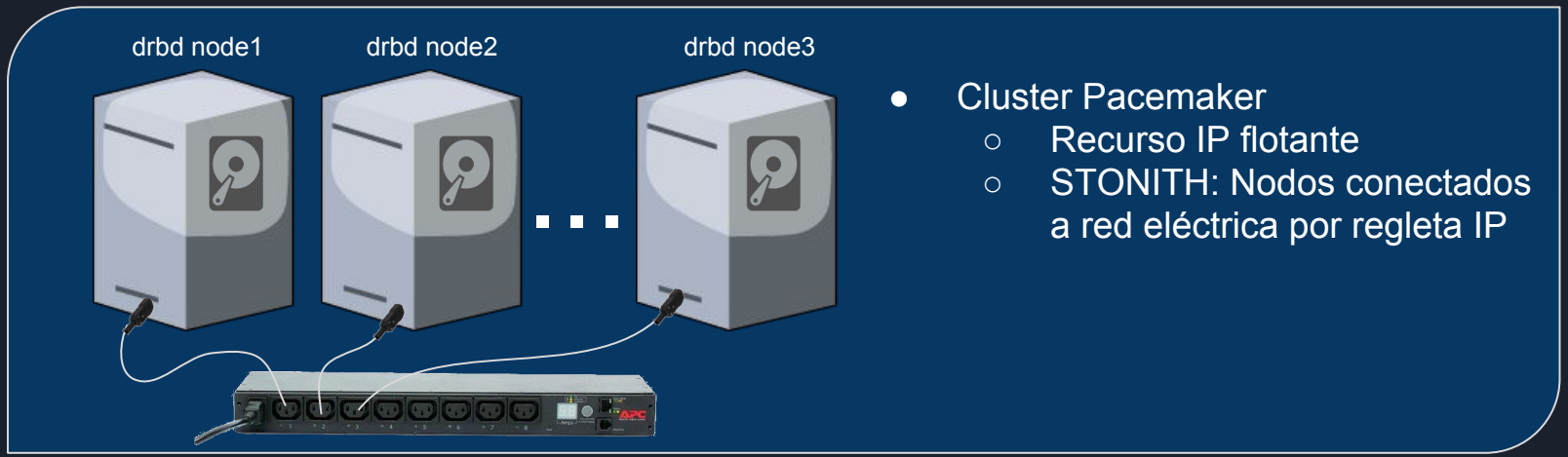
- Creamos recursos propios:
  - Para write caché

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

Alta Disponibilidad: Diagrama

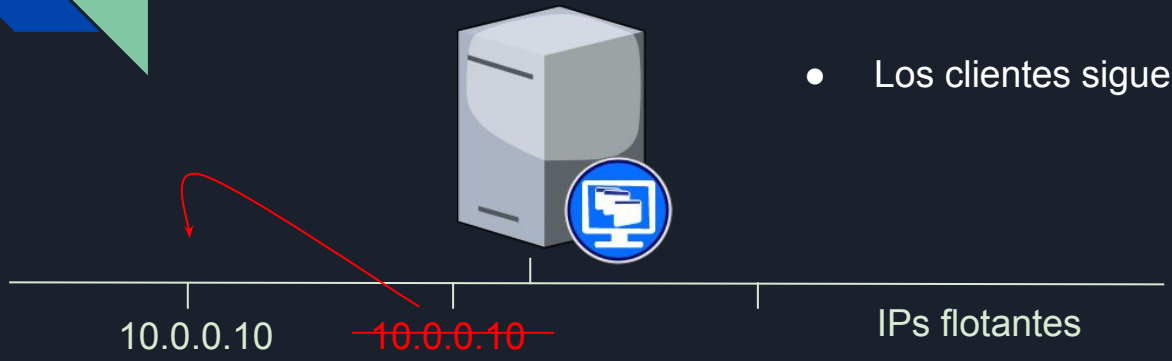


- Clientes montan almacenamiento desde ip flotante.

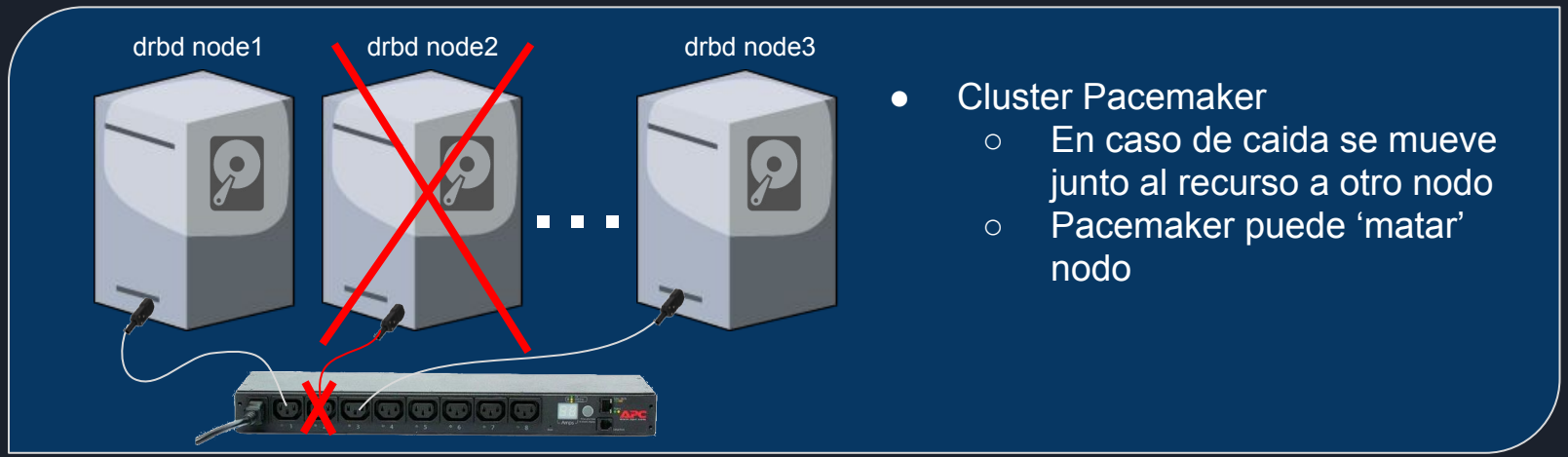


- Cluster Pacemaker
  - Recurso IP flotante
  - STONITH: Nodos conectados a red eléctrica por regleta IP

Alta Disponibilidad: Diagrama

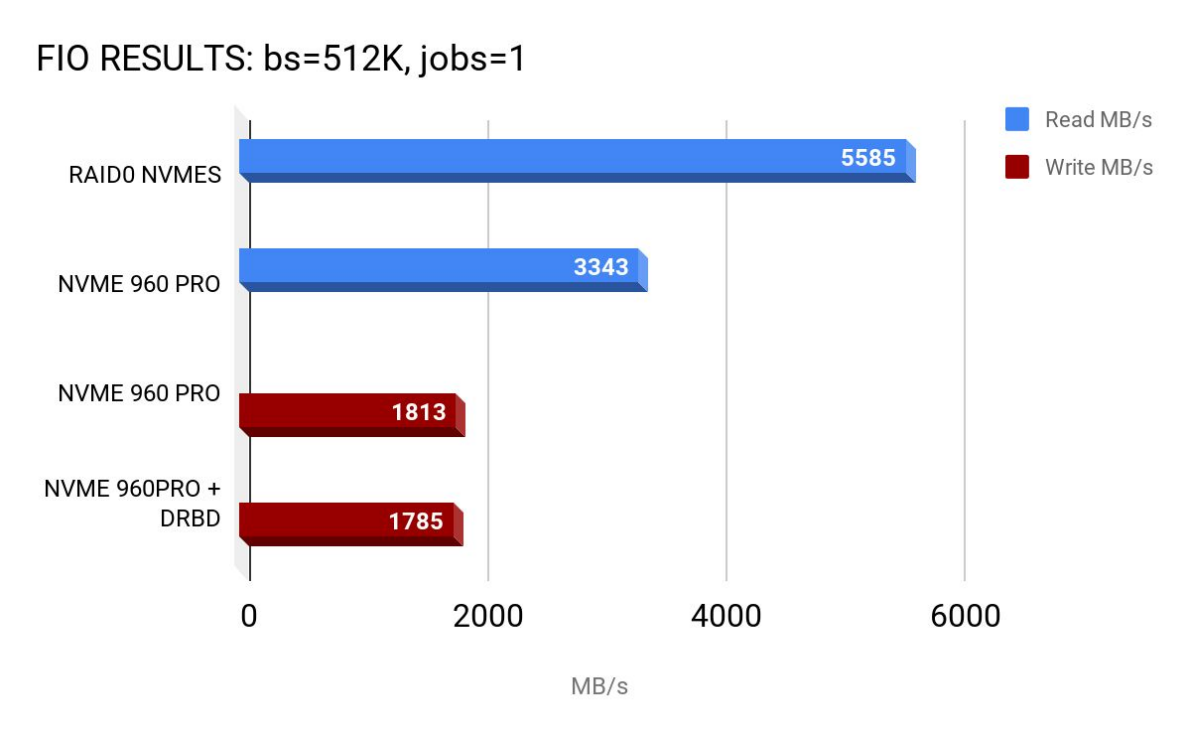


- Los clientes siguen trabajando



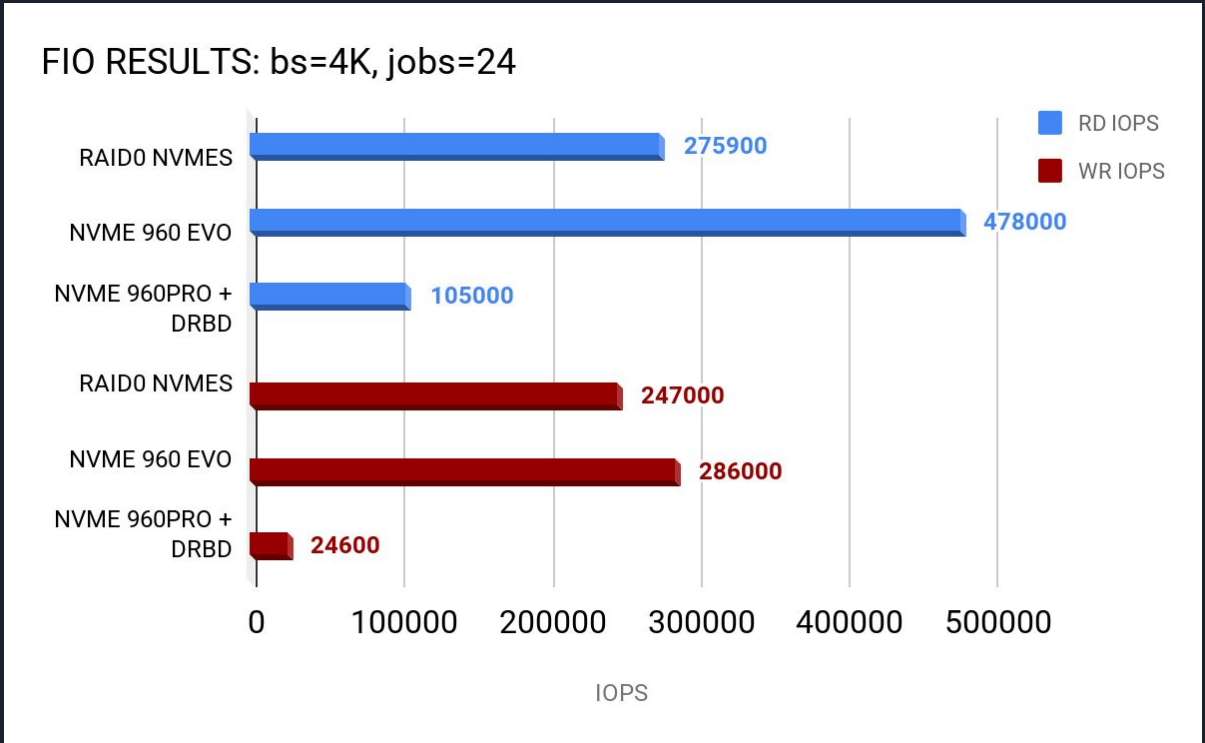
- Cluster Pacemaker
  - En caso de caida se mueve junto al recurso a otro nodo
  - Pacemaker puede ‘matar’ nodo

Resultados FIOS



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

Resultados FIOS



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

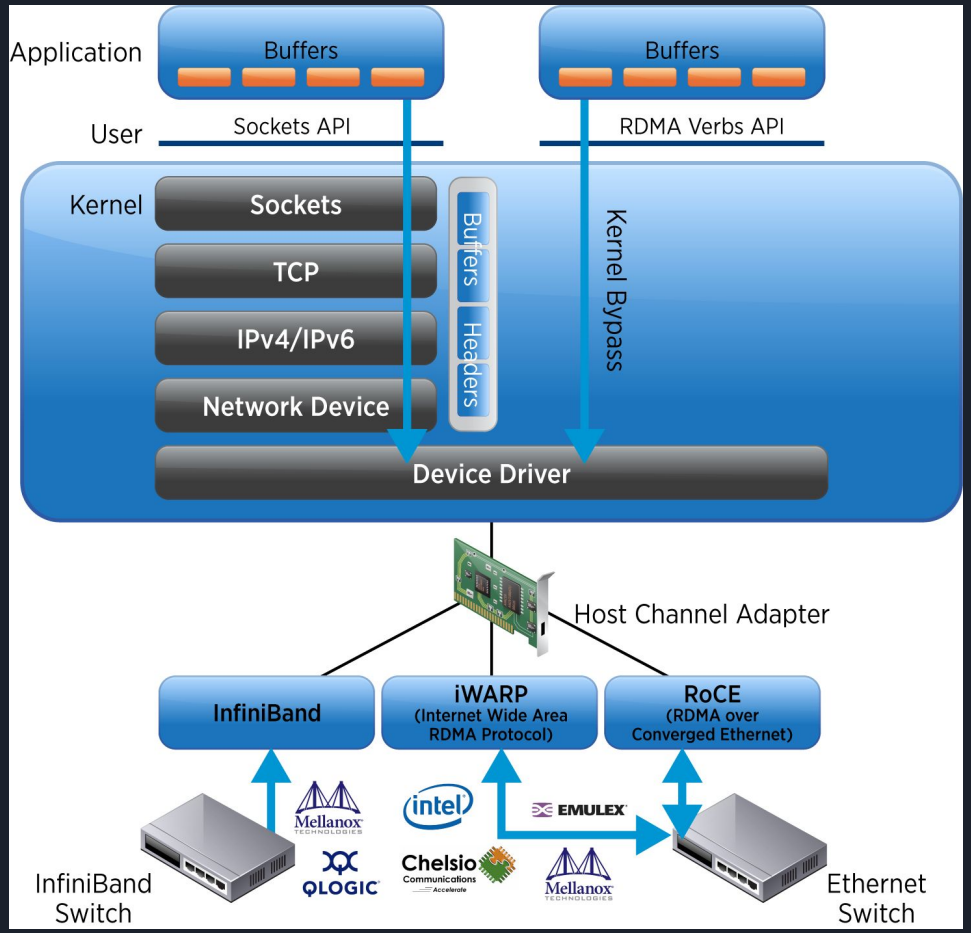
```
net.ipv4.tcp_rmem = 4096 87380 67108864
```

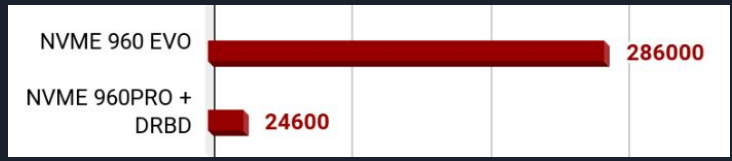
Mapeo estático o dinámico de IRQs para balancear cargas entre las CPU disponibles

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

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

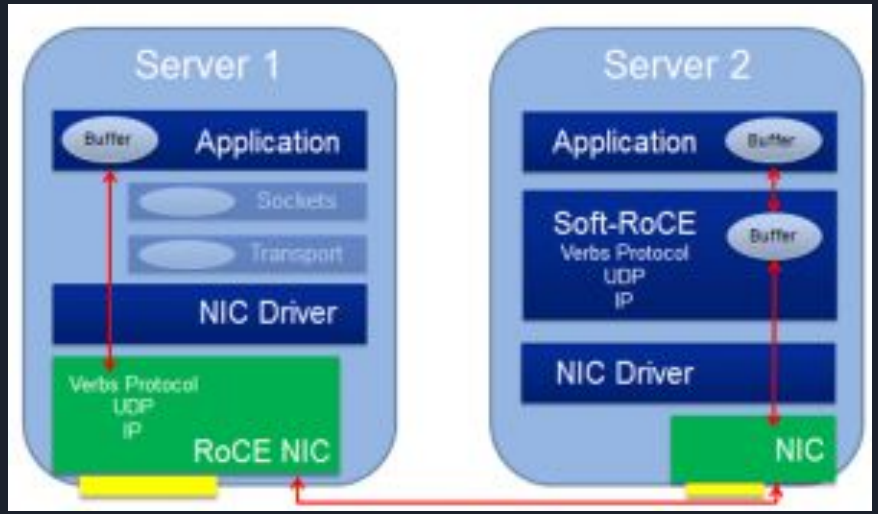
RDMA (remote direct memory access)



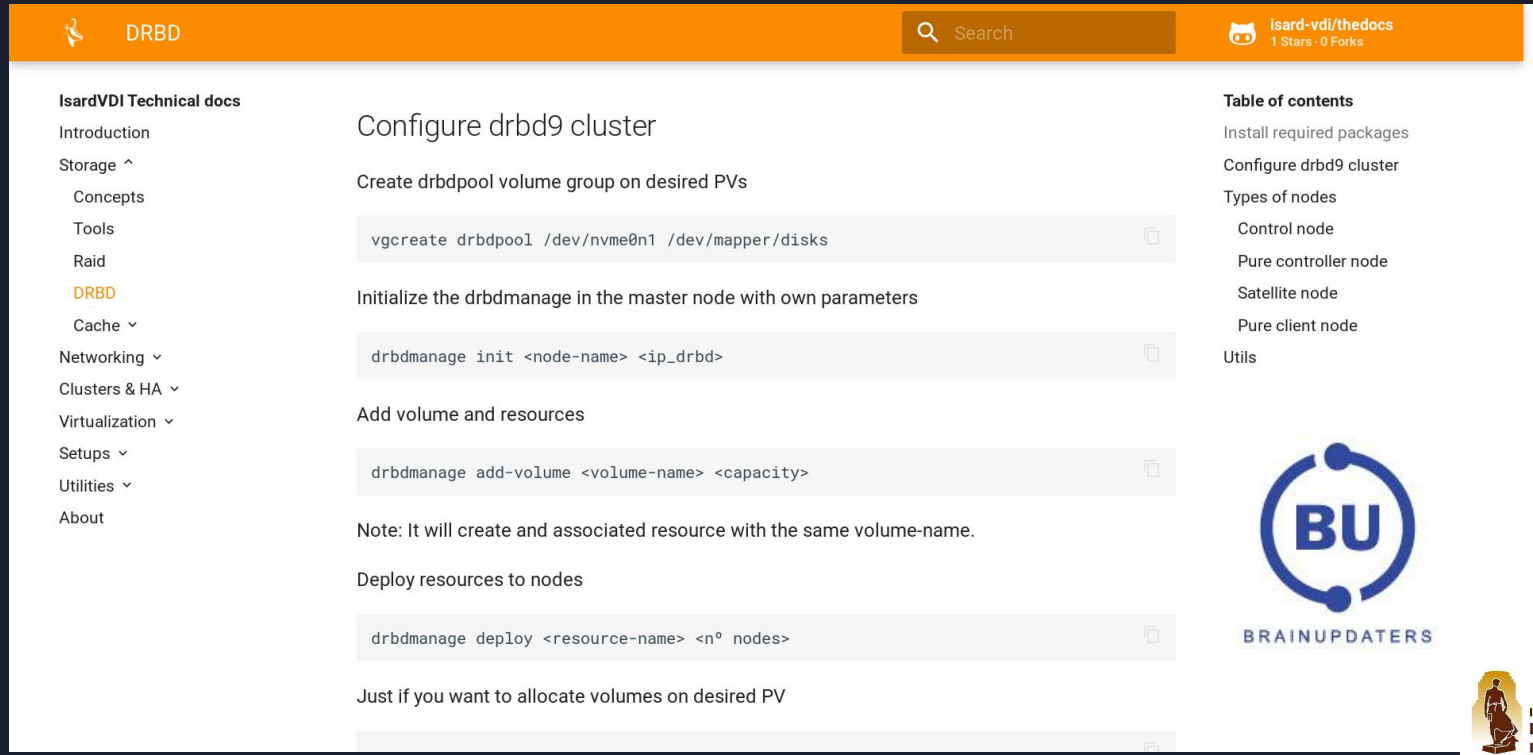


RoCE: RDMA over ethernet:

- Hardware RoCE
- Soft RoCE (`modprobe rdma_rxe`)



Documentación en techdocs.escoladeltreball.org



INSTITUT ESCOLA DEL TREBALL DE BARCELONA

XTEC Xarxa Telemàtica Educativa de Catalunya

Gracias por la atención

Guías de instalación, configuraciones, tests...

http://techdocs.escoladeltreball.org

https://github.com/isard-vdi

sysadmin@escoladeltreball.org