Jornadas Técnicas de RedIRIS 2015
Santa Cruz de Tenerife, 24-26 de noviembre de 2015

FÓRMULAS DE ALMACENAMIENTO BASADAS EN GLUSTERFS PARA EL SERVICIO DE CUENTAS DE USUARIO EN LABORATORIOS DOCENTES CON SOFTWARE LIBRE

Omar Aurelio Walid, Gabriel Martín, Héctor Bedoya, David Fernández

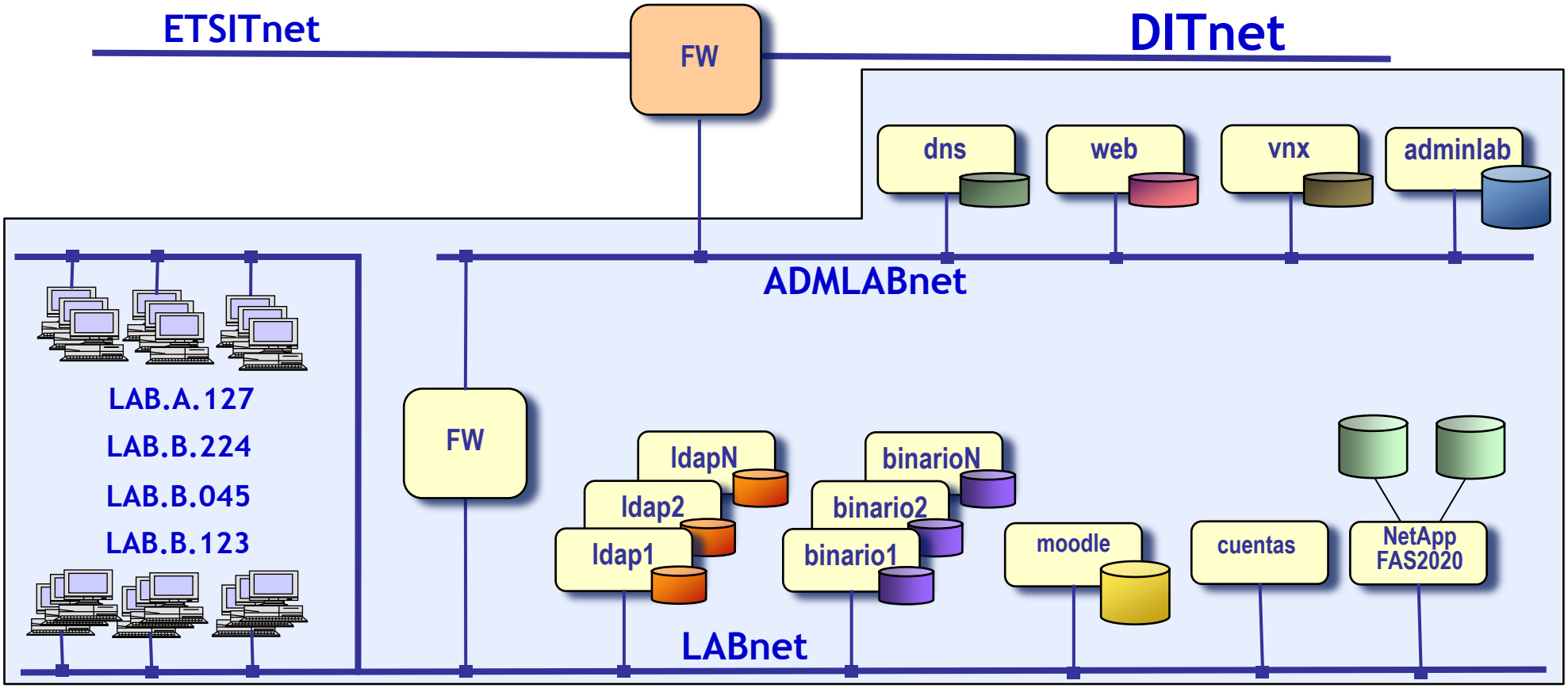
dit UPM

mailto: omar@dit.upm.es
Centro de Cálculo
Dpto. Ingeniería de Sistemas Telemáticos
ETSIT-UPM

Contenido

- Problemática de los laboratorios docentes
- Problemática del almacenamiento
- ¿Qué hemos usado hasta ahora? ¿Porqué hemos cambiado de paradigma?
- Soluciones basadas en software libre y reparto de carga

Laboratorios Docentes del DIT



¿Qué hemos usado hasta ahora?

- NAS comercial de NetApp (FAS2020)

Buen resultado, pero:

- Rendimiento: máximo de 60 MB/s
- Coste: mantenimiento $$$
- Funcionalidad: dependiente de $
- Repuestos de la marca ($$)

¿Porqué hemos cambiado de paradigma?

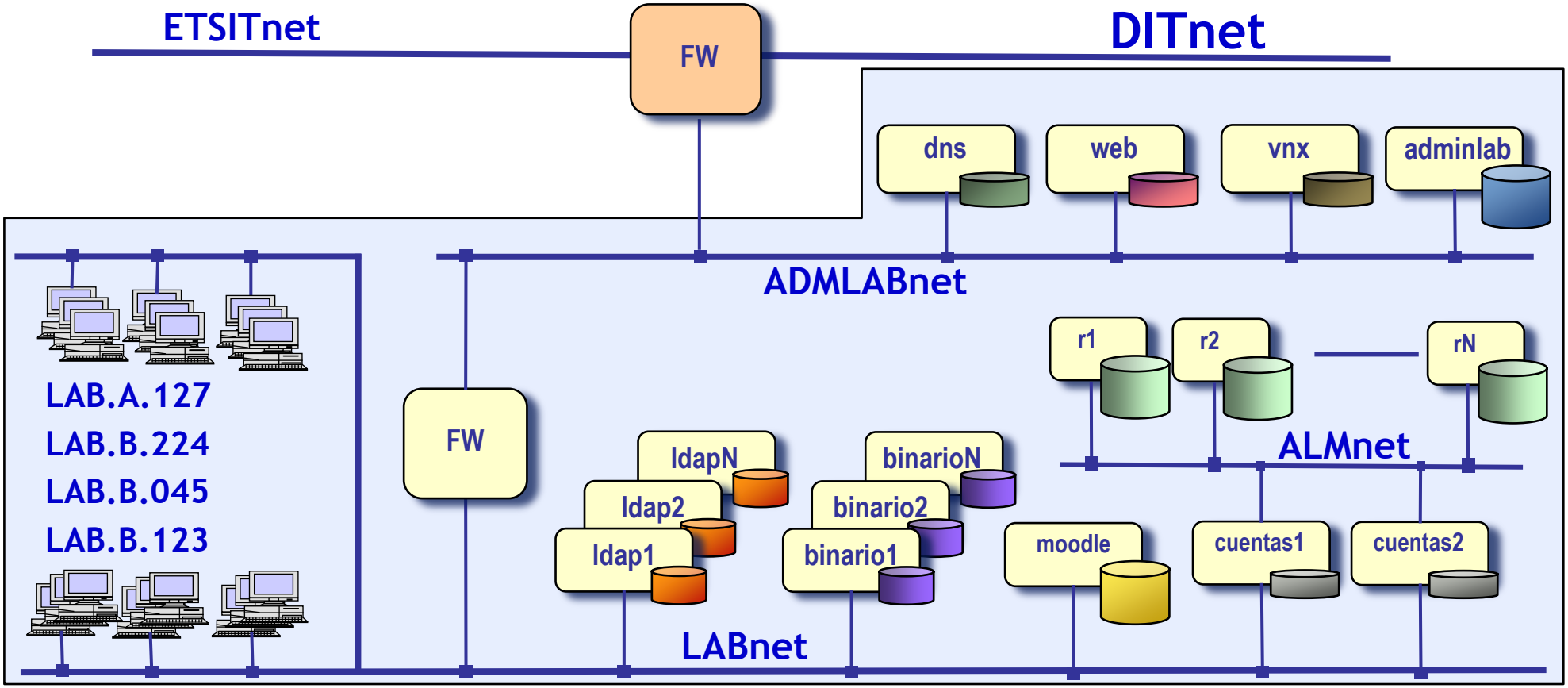
¿Herramientas? Software libre

Linux

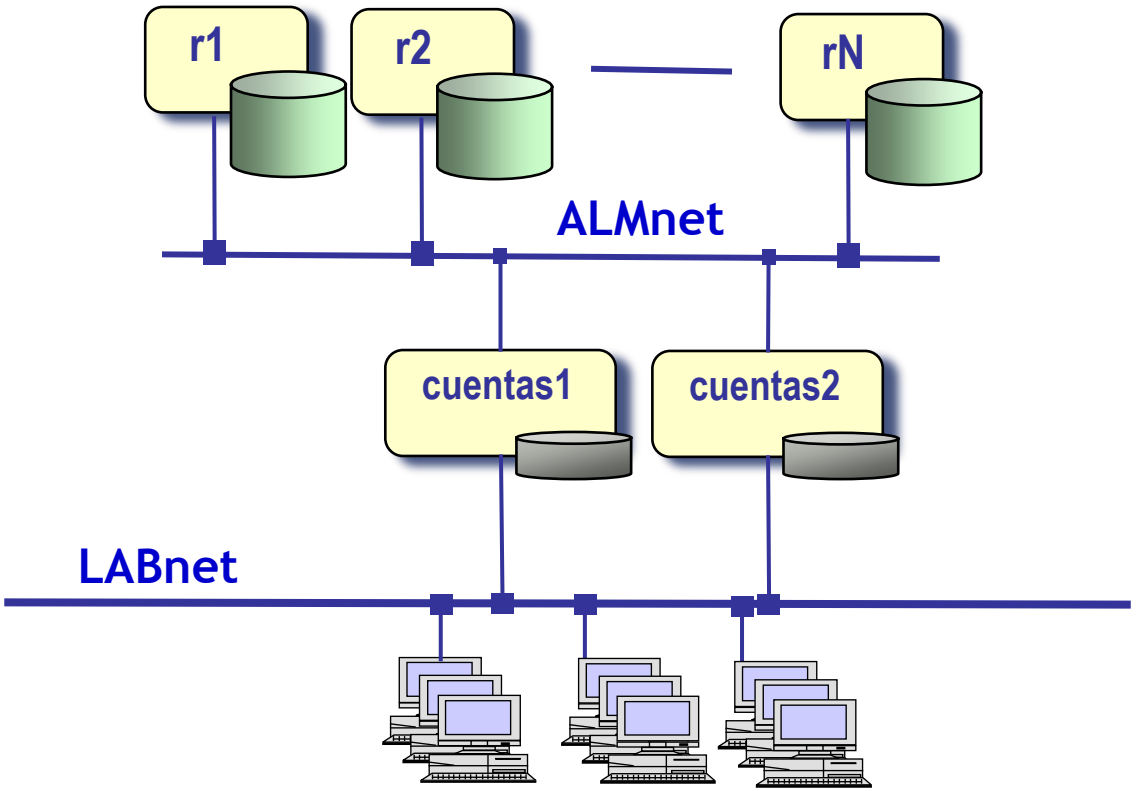
ZFS

ubuntu

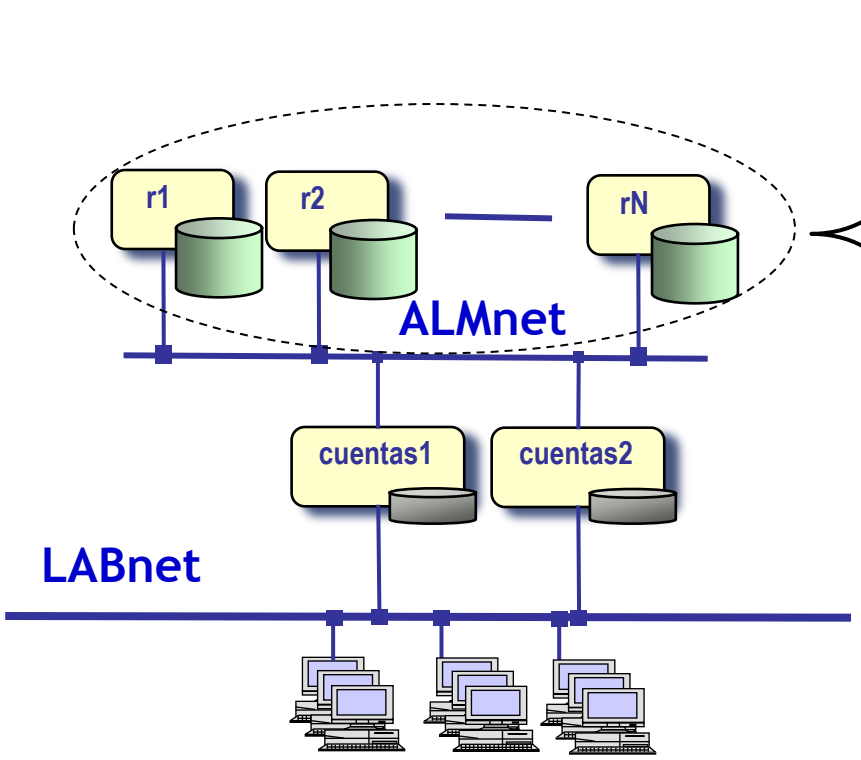
Nuevos laboratorios docentes del dit



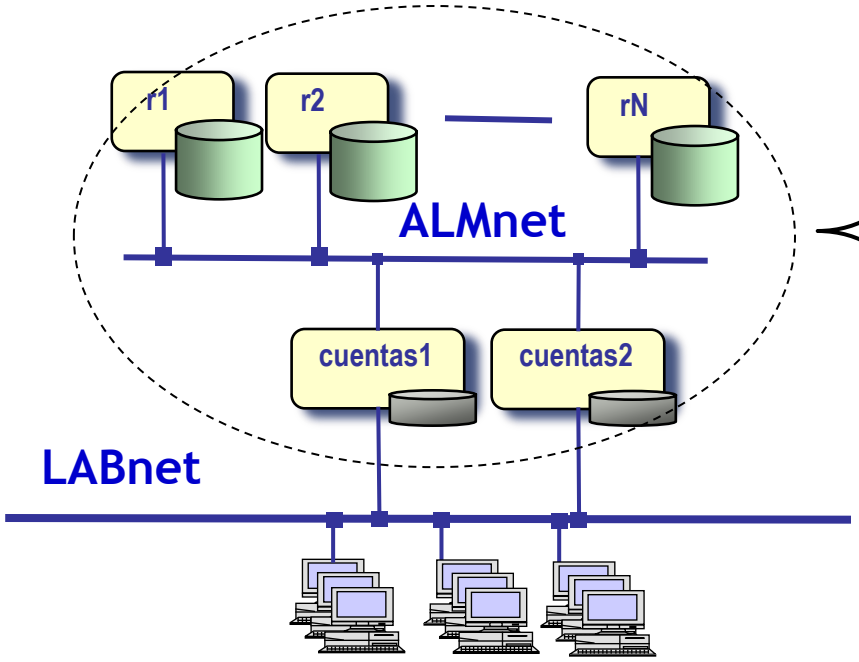
Nueva arquitectura de almacenamiento



Reutilización de recursos



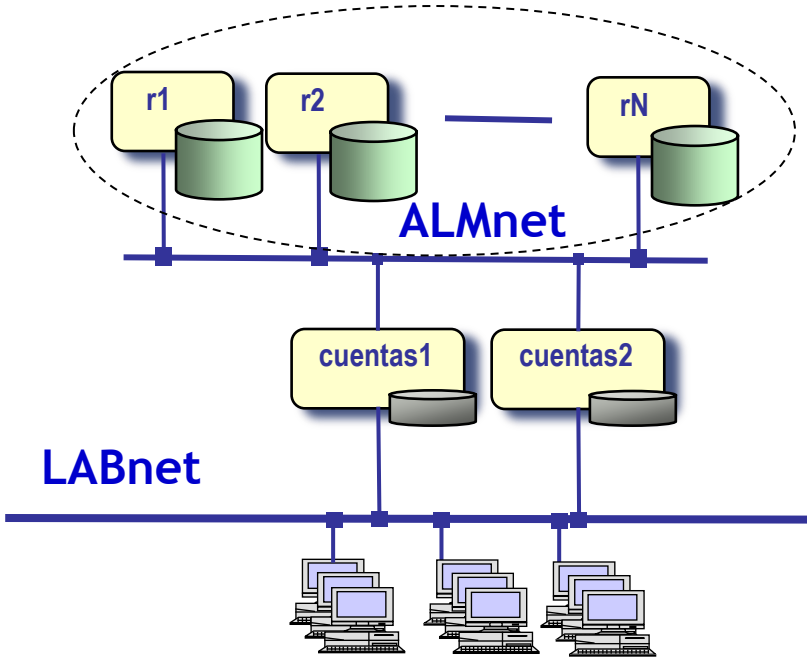
Herramientas básicas



Linux

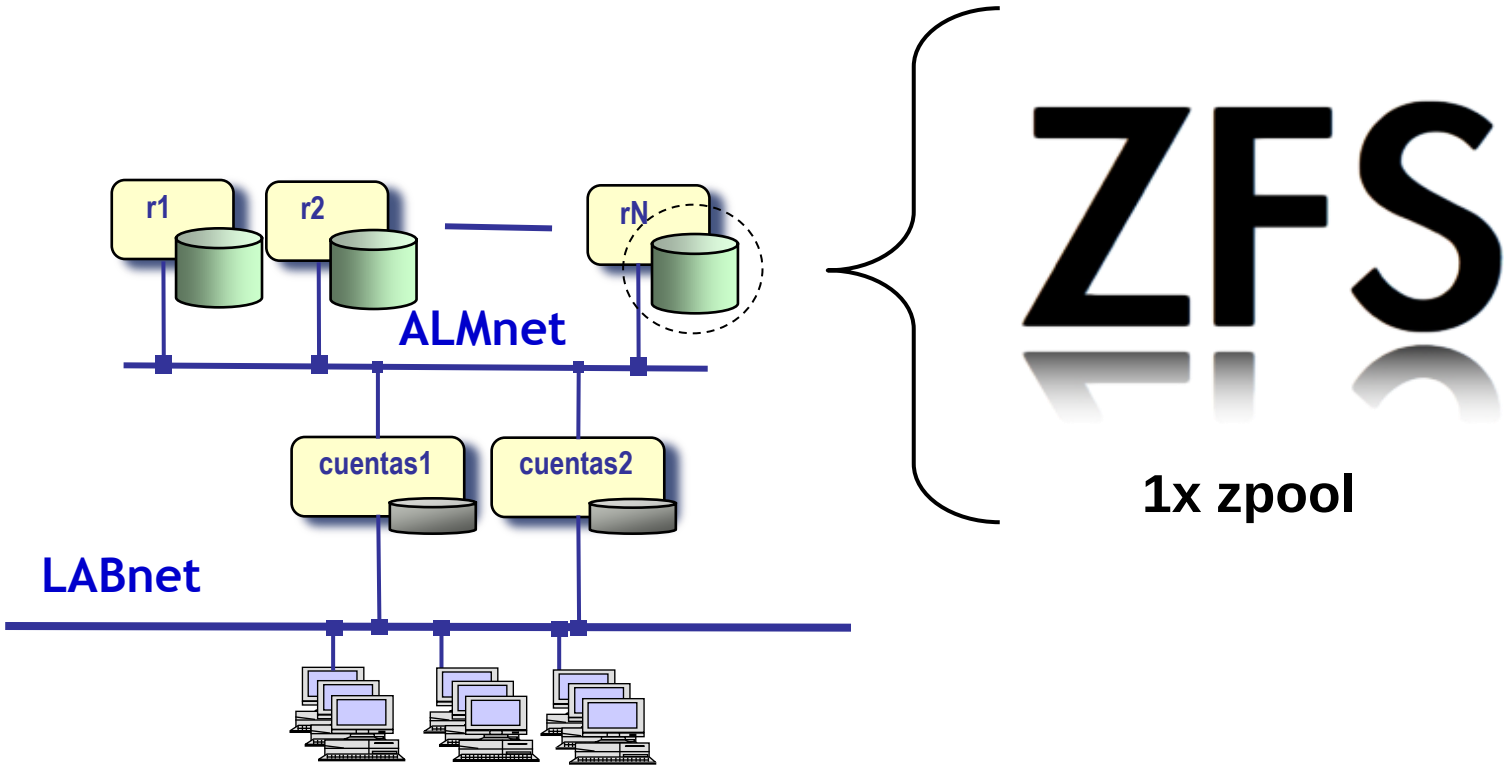
ubuntu

Arranque y S.O. en USB

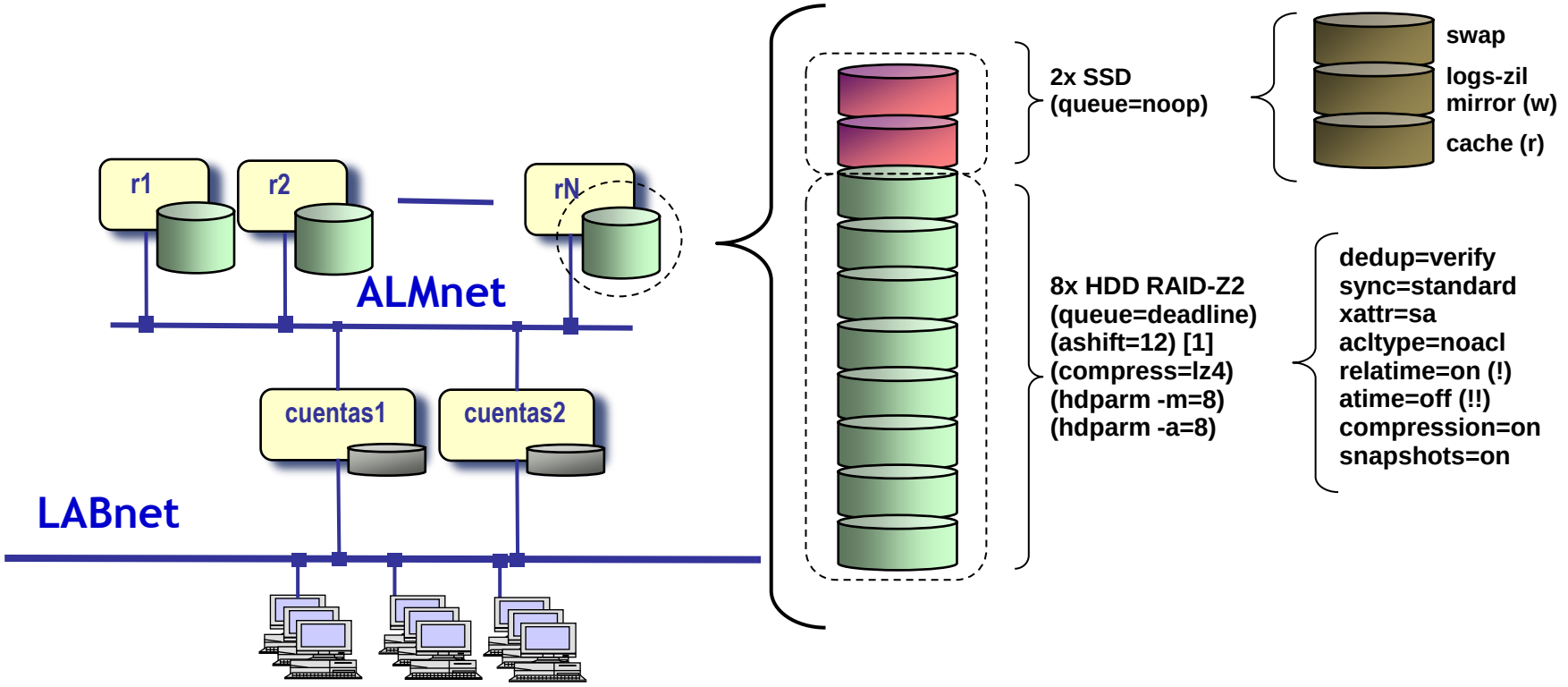


root mount=relatime
rsyslog

Almacenamiento en ZFS

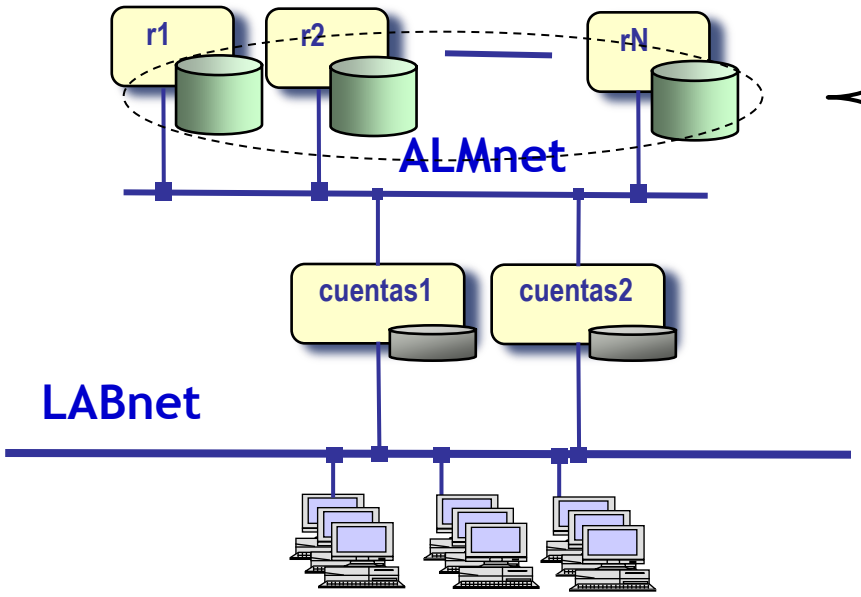


Configuración de ZFS



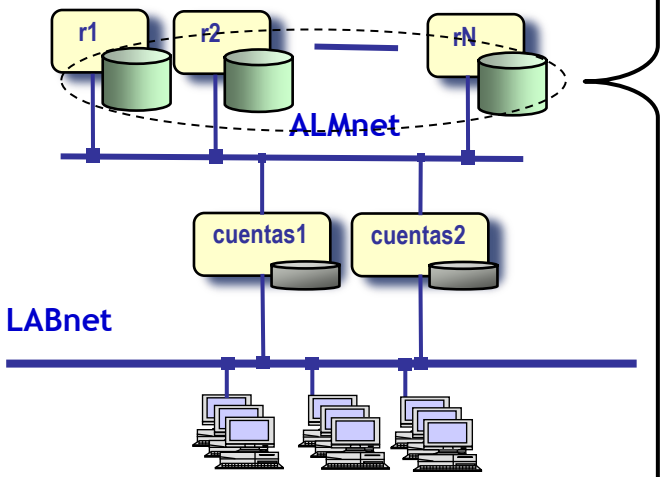
[1] http://louwrentius.com/zfs-performance-and-capacity-impact-of-ashift9-on-4k-sector-drives.html

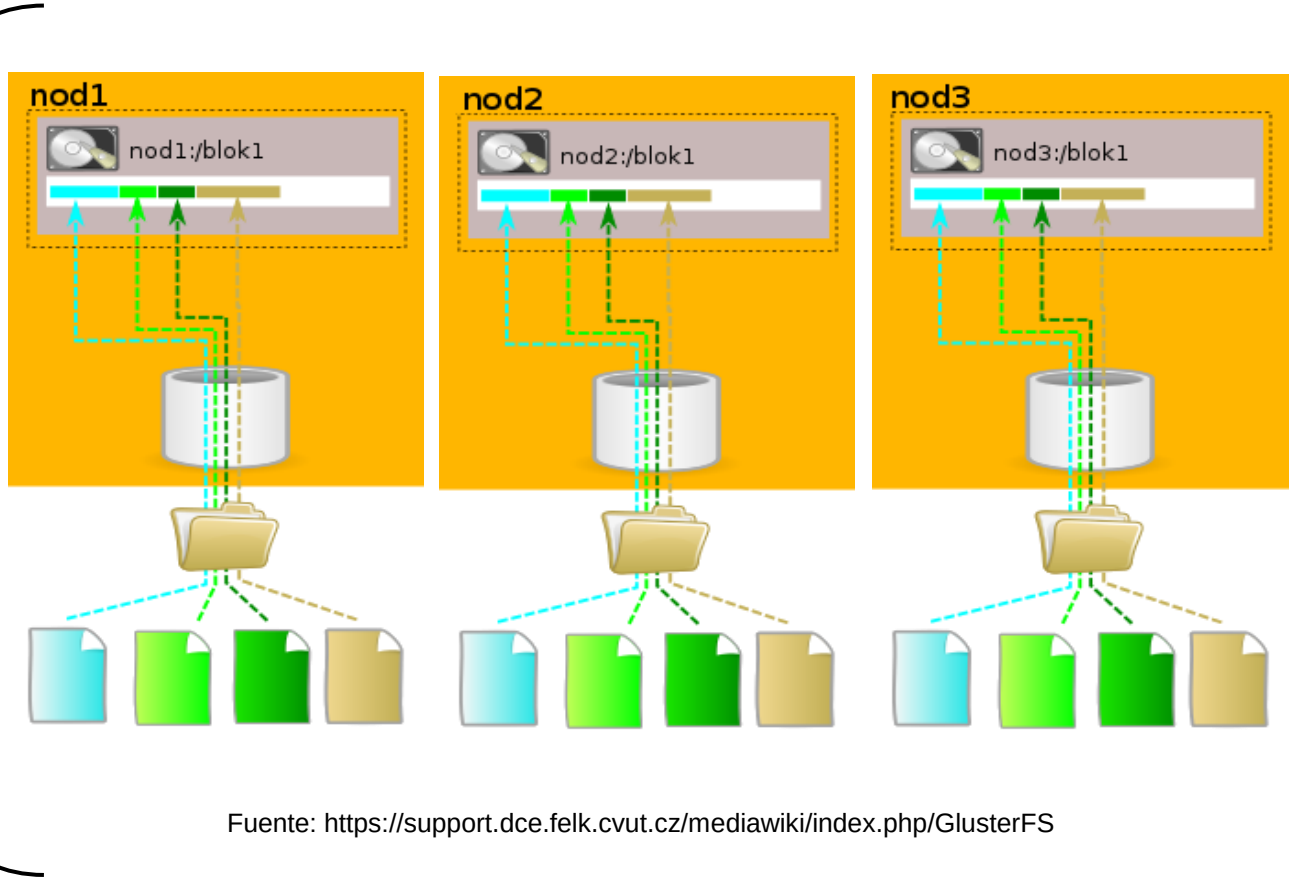
Cluster de almacenamiento con glusterfs



X gluster vol

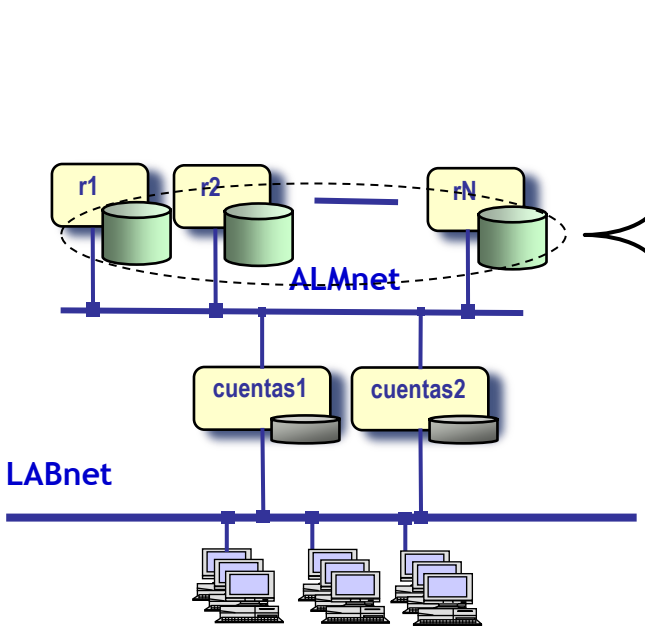
Glusterfs modo standalone

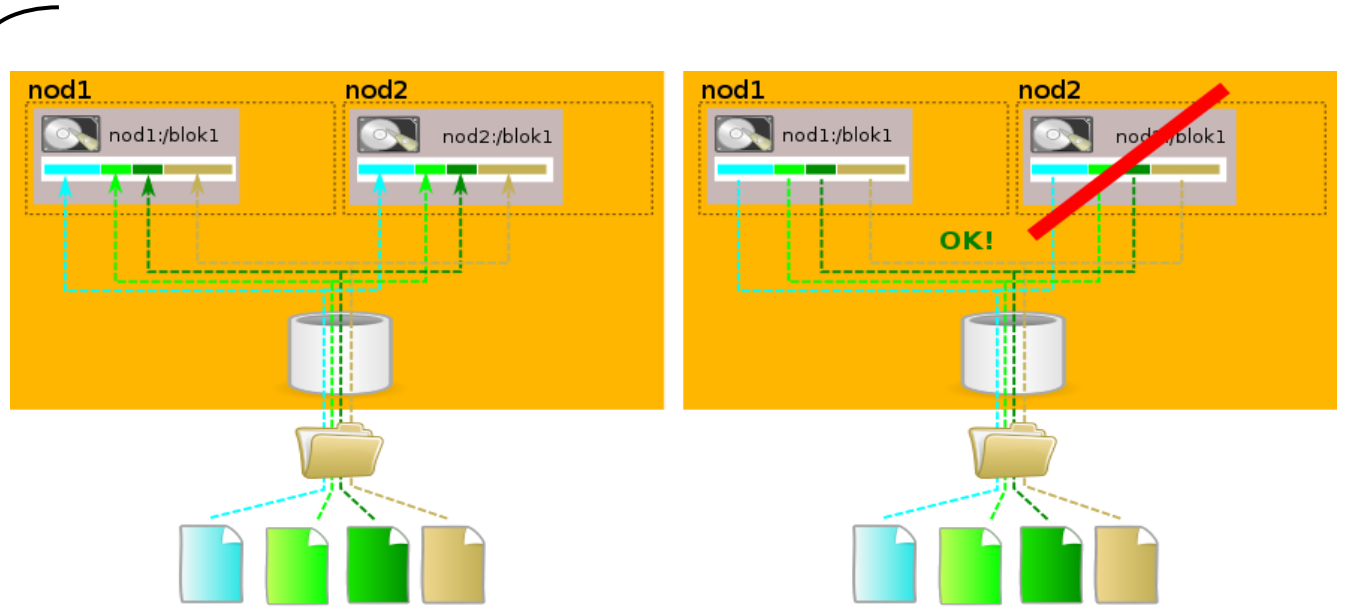




Fuente: https://support.dce.felk.cvut.cz/mediawiki/index.php/GlusterFS

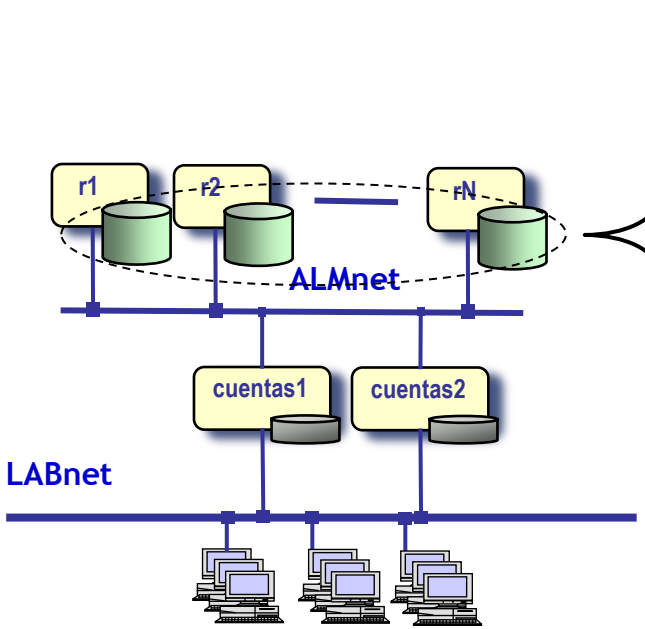
Glusterfs modo réplica

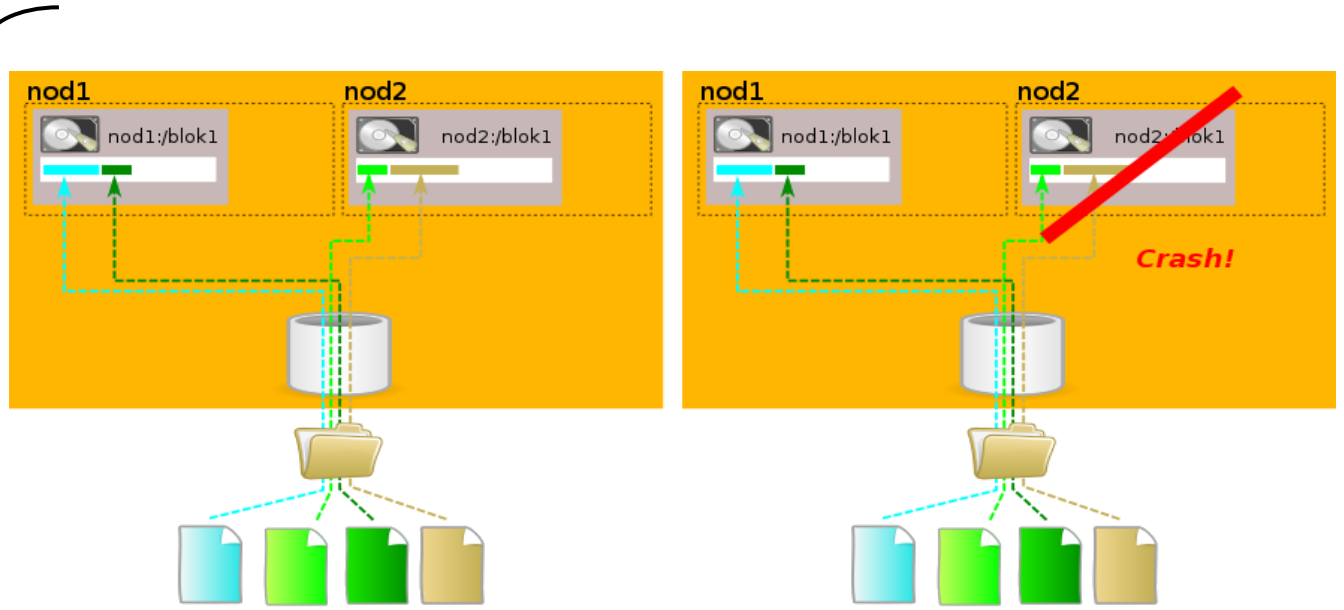




Fuente: https://support.dce.felk.cvut.cz/mediawiki/index.php/GlusterFS

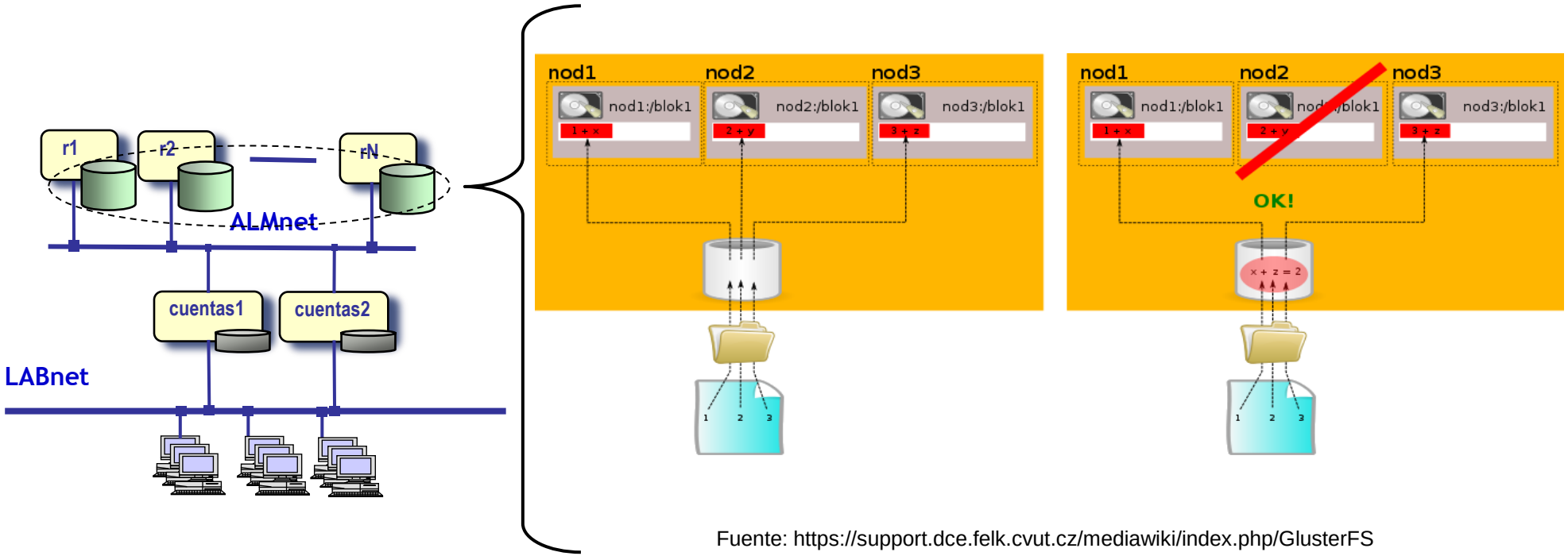
Glusterfs modo distribuido





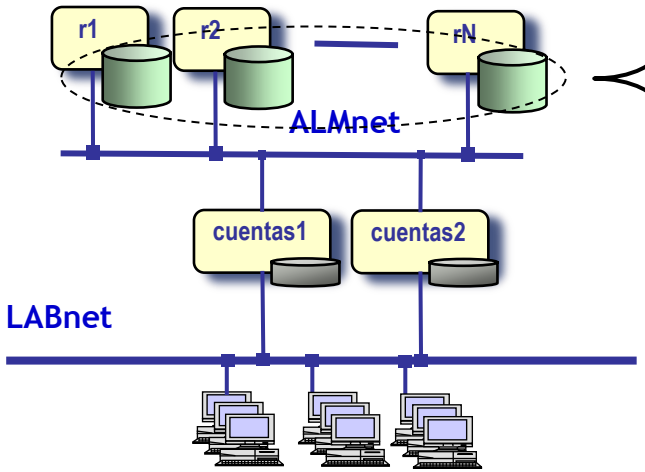
Fuente: https://support.dce.felk.cvut.cz/mediawiki/index.php/GlusterFS

Glusterfs modo disperso



Fuente: https://support.dce.felk.cvut.cz/mediawiki/index.php/GlusterFS

Configuración de glusterfs en el dit



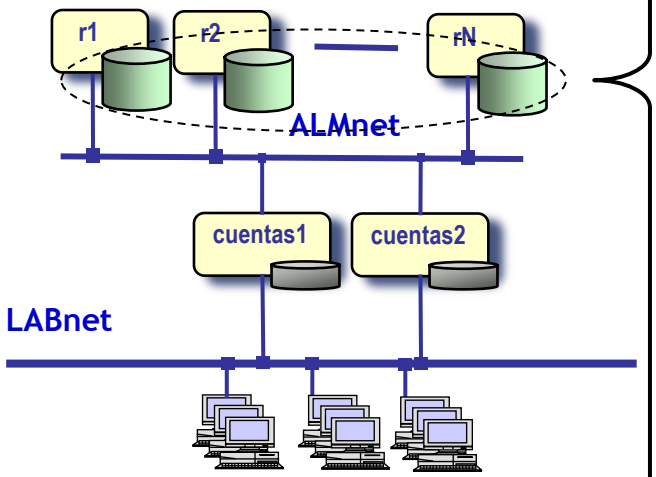
ppa:gluster/glusterfs-3.7 (versión 3.7.4)

on:
- allow-insecure
- auth.allow *
- diagnostics.dump-fd-stats
- performance.cache-size 256MB
- performance.client-io-threads
- server.event-threads 8
- client.event-threads 8
- server.outstanding-rpc-limit 128
- readdir-ahead
- readdir-optimize
- features.cache-invalidation
- performance.cache-max-file-size 128KB

off / disable:
- quota
- quota.deem-statfs
- nfs.export-dirs
- nfs.export-volumes
- nfs.register-with-portmap
- performance.open-behind
- features.quota-deem-statfs

X gluster vol

Volúmenes glusterfs resultantes



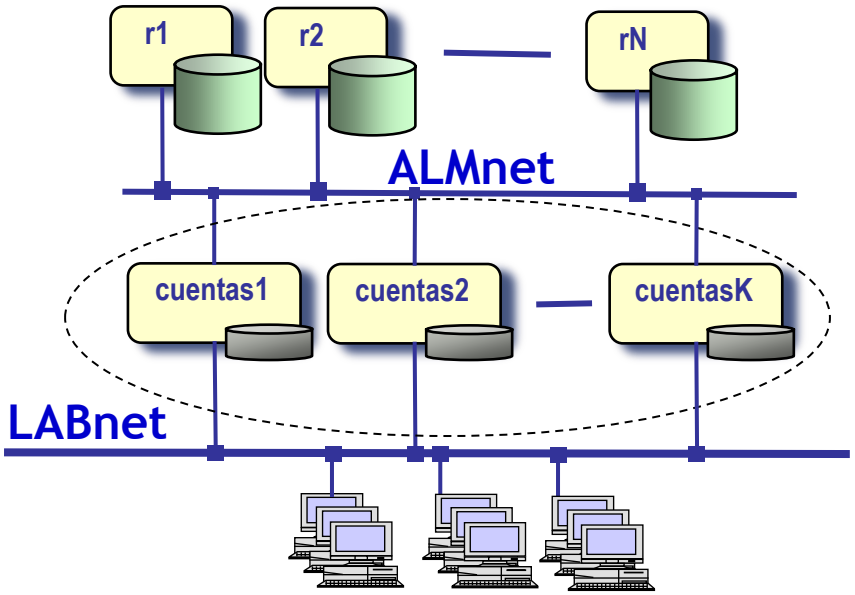
DIT (N=4): cluster de 4 servidores

4x home-0 (vol 1 brick modo standalone)
2x home-1 (vol 2 bricks modo réplica)
2x home-2 (vol 2 bricks modo distribuido)
1x home-3 (vol 4 bricks modo distribuido)
1x home-4 (vol 4 bricks modo disperso)
1x home-5 (vol 4 bricks modo réplica)

ZFS + glusterfs =
- Escalabilidad
- Potencia
- Rendimiento
- Replicación

11x gluster vol

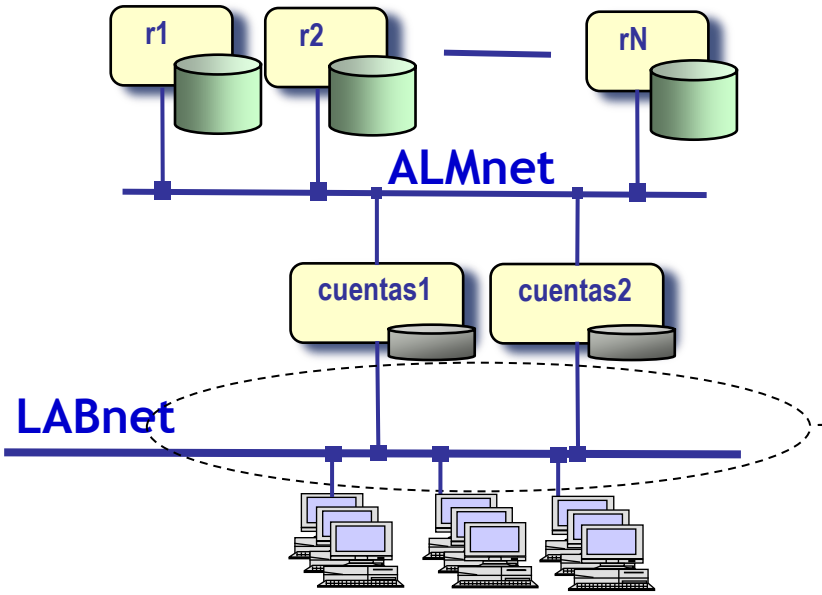
Re-exportación del almacenamiento



- Cliente: glusterfs
  Servidor: NFS/CIFS
  VM
  - aislamiento
  - escalabilidad
  - fiabilidad
- Creación de cuentas
  Tareas periódicas
  Autenticación por LDAPS
  Gestión estadística y de seguridad
- DIT (K=2): cluster de 2 servidores

5x HOMES

Acceso al almacenamiento



Montaje de HOME desde menú { NFS, CIFS }

Caché local a disco - cachefilesd (fsc)
Opciones: relatime, nolock, wsize, rsize

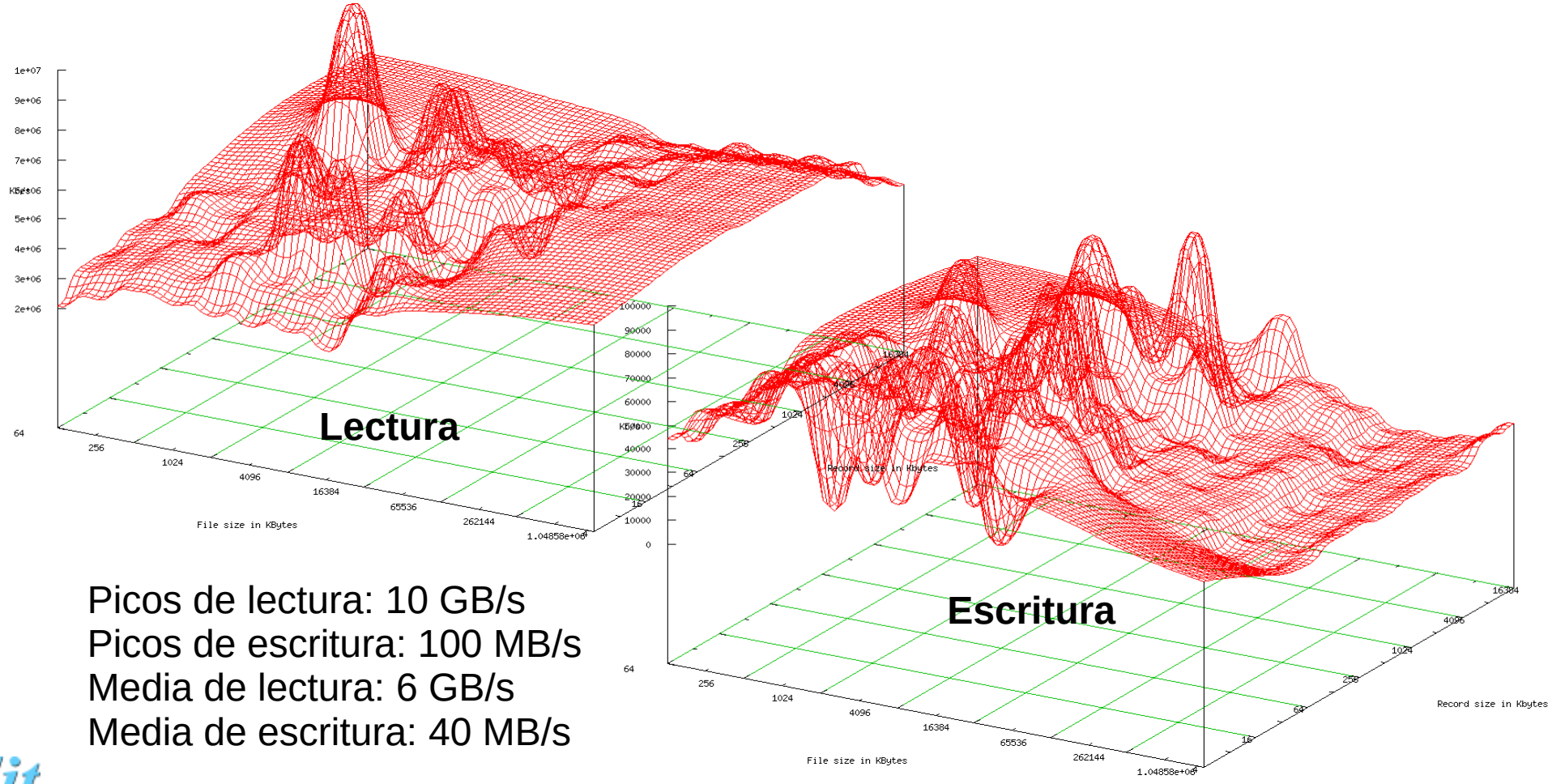
nofsync[1] en LD_PRELOAD

Montaje de otros HOMES en paralelo

} picos 1000 req/s

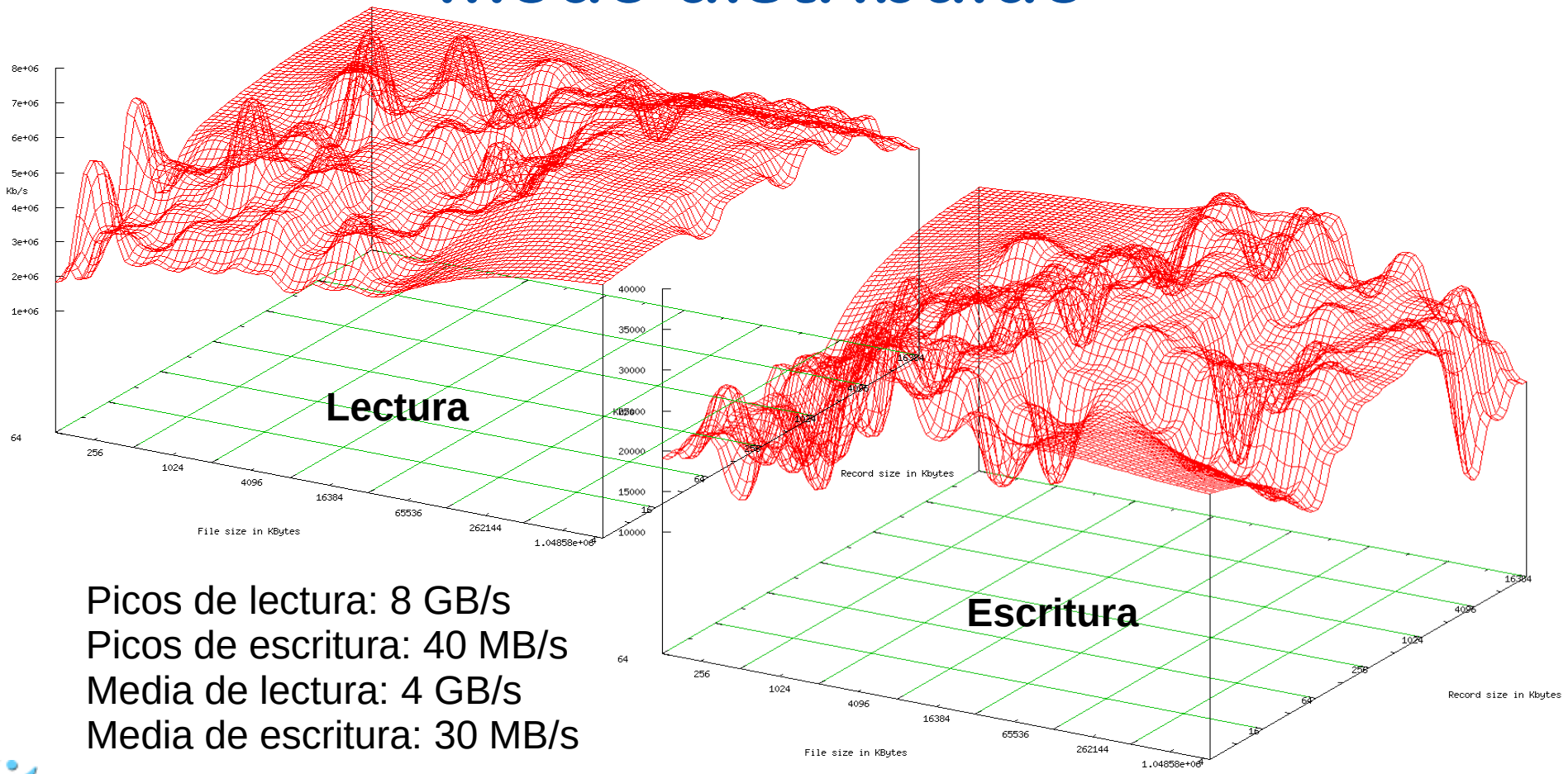
[1] http://ubuntuforums.org/archive/index.php/t-1103926.html

Pruebas NAS NetApp



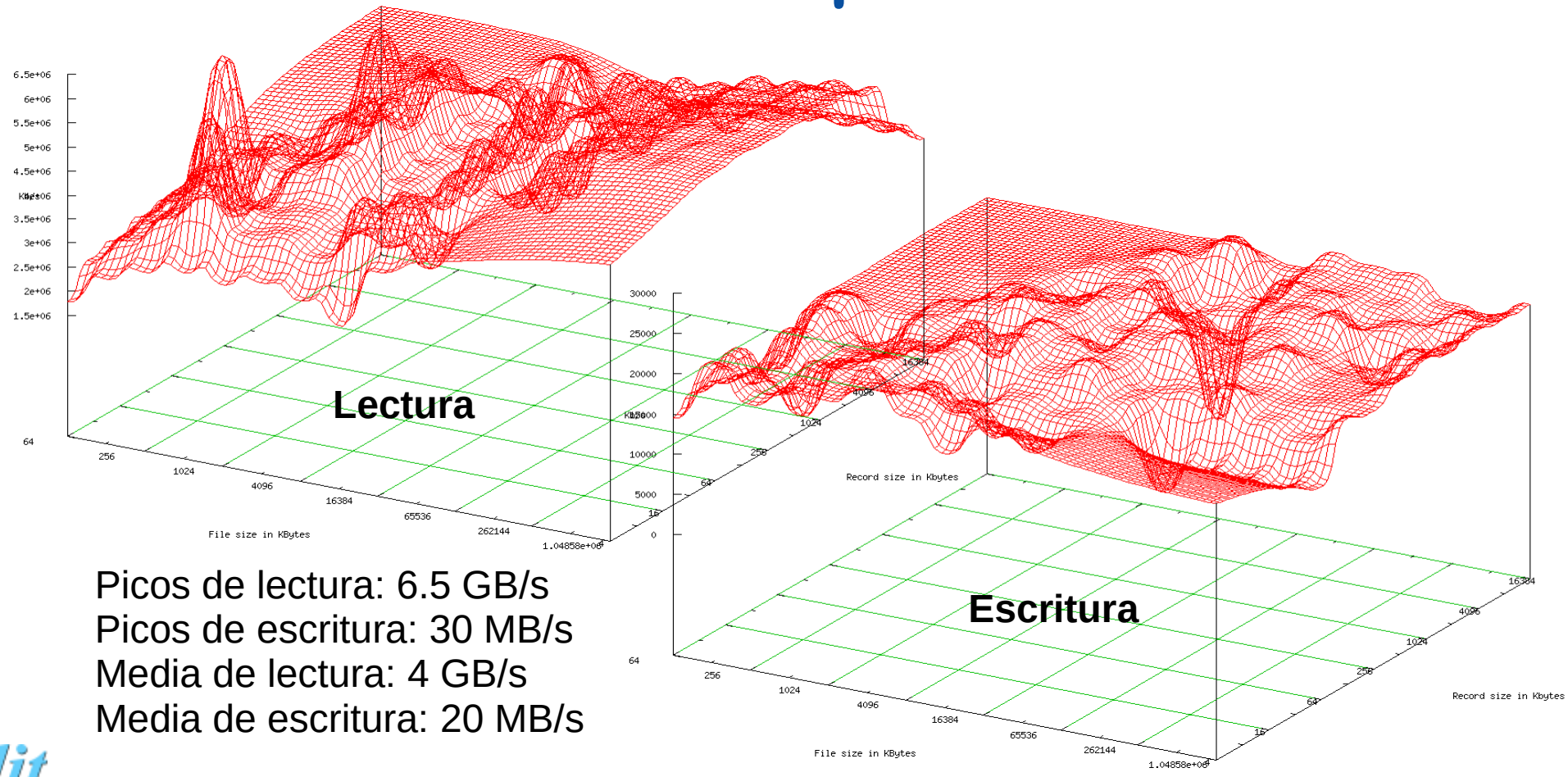
Picos de lectura: 10 GB/s
Picos de escritura: 100 MB/s
Media de lectura: 6 GB/s
Media de escritura: 40 MB/s

Pruebas nfs-kernel sobre glusterfs modo distribuido



Picos de lectura: 8 GB/s
Picos de escritura: 40 MB/s
Media de lectura: 4 GB/s
Media de escritura: 30 MB/s

Pruebas nfs-kernel sobre glusterfs modo disperso



Picos de lectura: 6.5 GB/s
Picos de escritura: 30 MB/s
Media de lectura: 4 GB/s
Media de escritura: 20 MB/s

Conclusiones de las pruebas

- Pruebas reales sobre el laboratorio:
  - nfs-kernel sobre gluster en modo distribuido
  - arranque simultáneo de >150 ordenadores
  - usuarios independientes
  - >10 aplicaciones entorno unix (sistema gráfico, eclipse, firefox, chrome, 2 terminales de texto, 1 documento pdf, etc)
    - Tiempo ahora: 15 minutos (antes > 30 minutos!!)
- El modo disperso es una solución a la que queremos llegar, pero de momento, y en nuestro entorno, el rendimiento del modo distribuido es del orden de 3 veces mejor.
- Datos: iozone -ac -g 1G; Gráficos: iozone_visualizer.pl; Tratamiento: gimp.

Problemas, 'habelos hailos'

- Self-heal automático de glusterfs: de momento, mejor en off.
- Snapshots de glusterfs: de momento, mejor no usarlo.

  ZFS ofrece snapshots de alta calidad (que gestionan las diferencias y ofrecen mucho mejor rendimiento, más info: zfs-auto-snapshot, zfs send/receive)
- Cuotas de glusterfs: de momento, mejor en off.

  ZFS ofrece mejor gestión de las cuotas (mejor rendimiento) y cuotas no solo a nivel de usuario, sino también a nivel de grupo.
- Listados de directorios, control de espacio usado y administración/gestión de usuarios, mejor desde ZFS si el volumen es distribuido (no si es disperso).
- Inodos de 64b para máquinas que tienen kernel de 32b. Hay solución, pero hay que tocar LD_PRELOAD: http://www.tcm.phy.cam.ac.uk/sw/inodes64.html
- nfs-kernel-server tiene un problema al copiar ficheros read-only (ej: git).

  Este problema no ocurre si se re-exporta con Samba/CIFS o nfs-ganesha.

Agradecimientos

Xavier Hernández (datalab.es). Desarrollador de GlusterFS (modo disperso).


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
BİYAN
SHUKRIA
TINGKI
GRAZIE
MEHRBANI
KOMAPSUMNIDA
MAAKE
PALDIES
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
BOLZÏN
MERCI
SPASSIBO
SNACHALHUYA
CHALTU
WABEEJA
MAITEKA
SPASIBO
DENKAUJA
SIKOMO
MINMONCHAR