CENTUM
CIEN AÑOS DE LA UNIVERSIDAD DE MURCIA
1915 | 2015

Implantación de ZFS para servicio de ficheros corporativo en la UMU

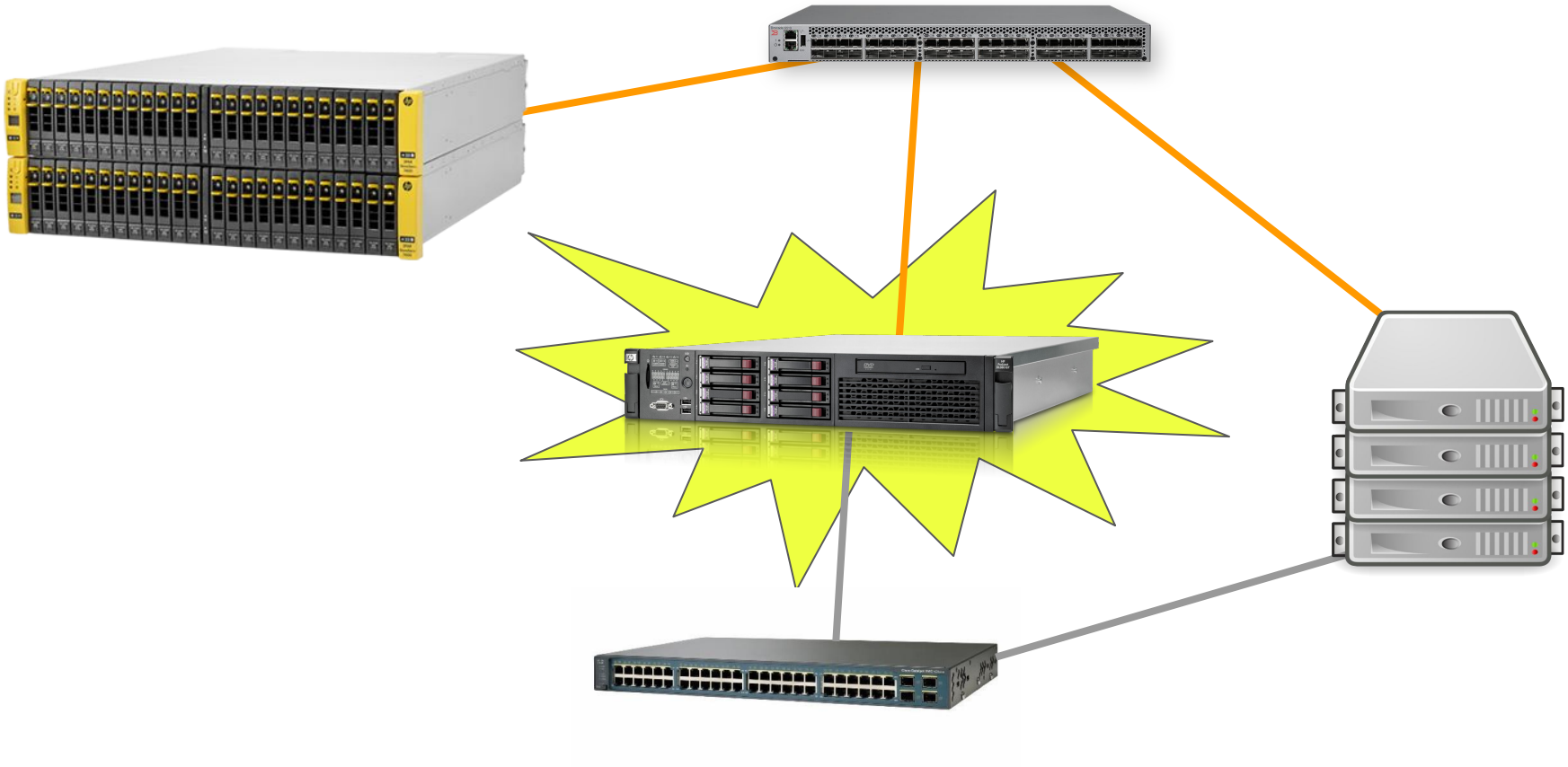
Pablo López Mozas - Universidad de Murcia
Jose Fco. Hidalgo Céspedes - Universidad de Murcia

UNIVERSIDAD DE MURCIA

Resumen de la charla

- Introducción al entorno de almacenamiento
- Decisión del producto elegido: ZFS
- Descripción de la nueva infraestructura NAS
- Estructura de ZFS
  - Pooles y recursos
  - Tareas habituales
  - Scrubs
  - Cuotas
  - Failover cluster
  - Snapshots y réplicas
- Monitorización
- Servicios alrededor de NFS
- Tareas futuras

Estructura de almacenamiento



Decisión sobre la arquitectura

hadoop HDFS

GlusterFS

Ext4 File System

vmware

CentOS

pNFS.com

ZFS ON LINUX

openfiler

FreeNAS™

LVM

¿Por qué ZFS?

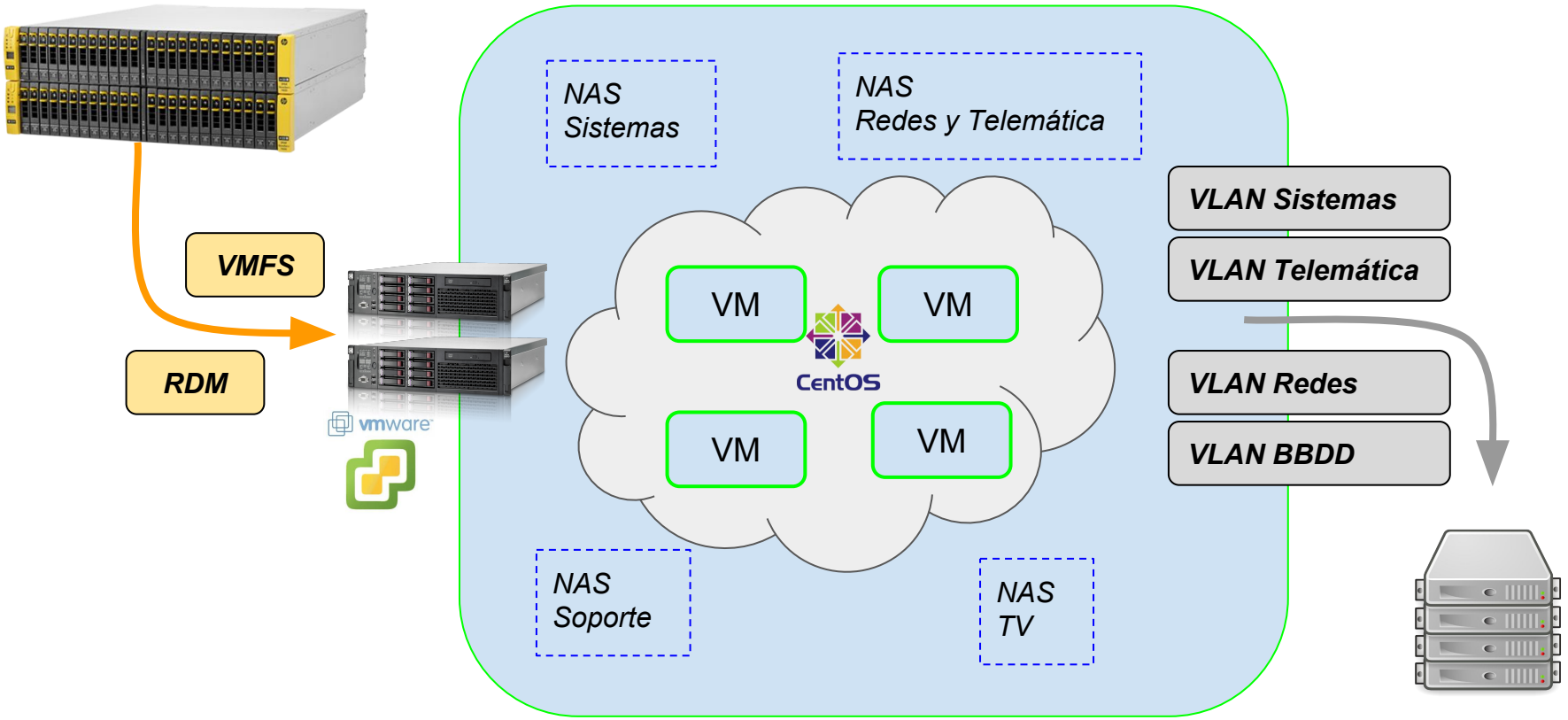
A favor de ZFS

- Software libre
- Conocimiento previo
- Características esperables
- Buenas críticas y recomendaciones

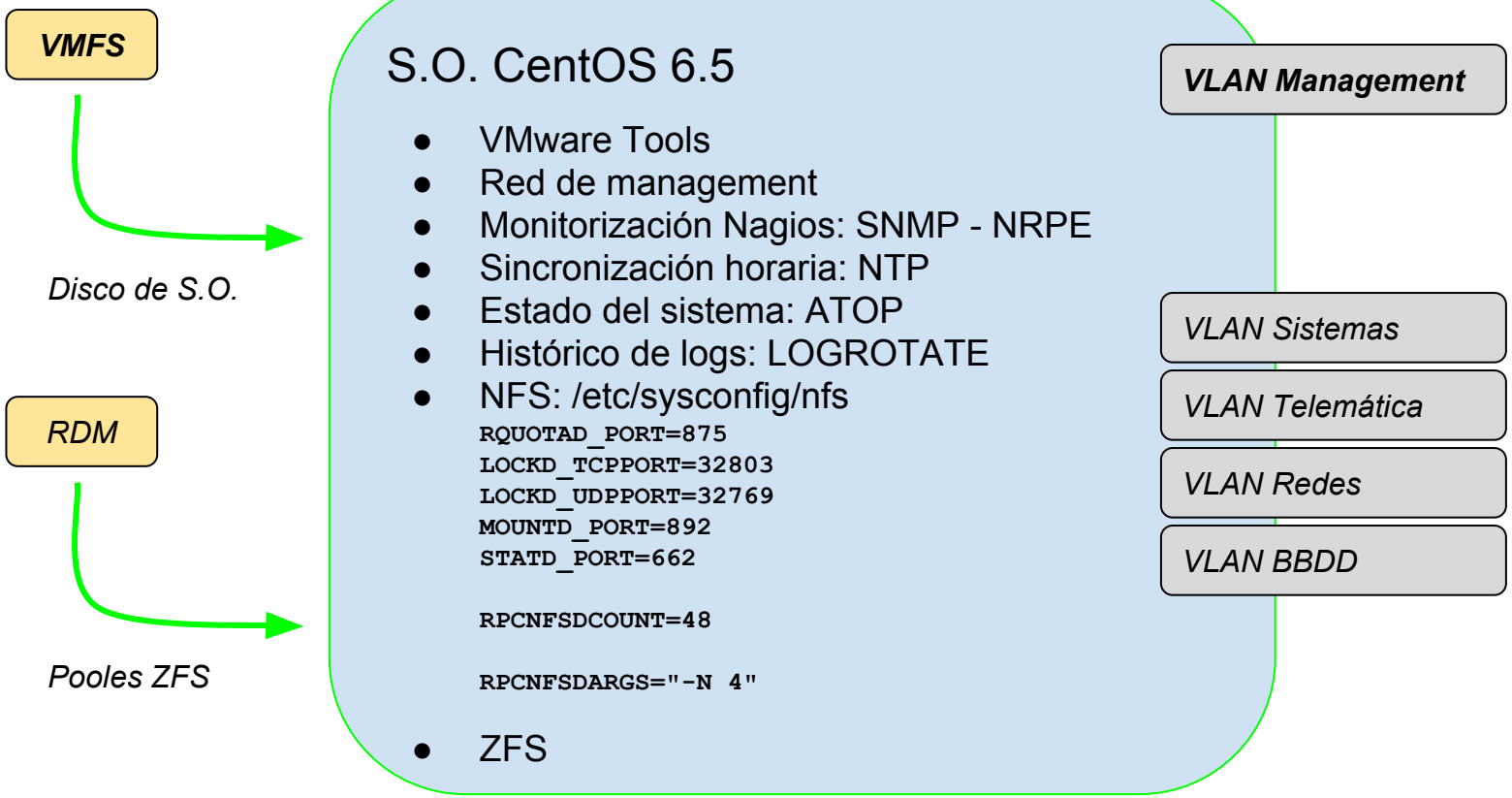
En contra de otros

- EXT4: antiguo y superado
- XFS: pocas características
- LVM: sólo particionado
- Red Hat Cluster: complejo, pocos FS
- Appliances: cerrados, funciones de pago

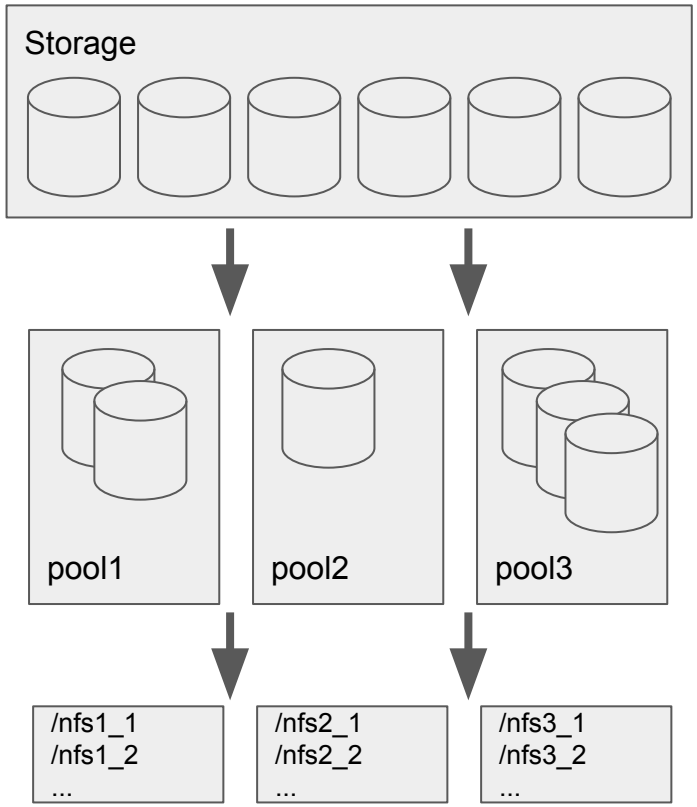
Descripción de la arquitectura



Sistema Operativo y Servicios



Estructura de pooles ZFS



- Un “filer” diferente para cada servicio: Sistemas, Telemática (RyT), Soporte-Novell, TV
- Pool como agrupación de uno o más volúmenes, mismo tipo de disco, y con un determinado fin

```
zpool create -f pool1 sdb sdc
zpool add -f pool1 sdj
```

- Los recursos se definen sobre los pooles

```
zfs create -o quota=200G -o mountpoint=/nfs1_1 pool1/nfs1_1
zfs set sharenfs=rw=155.54.208.17:155.54.208.25 pool1/nfs1_1
```

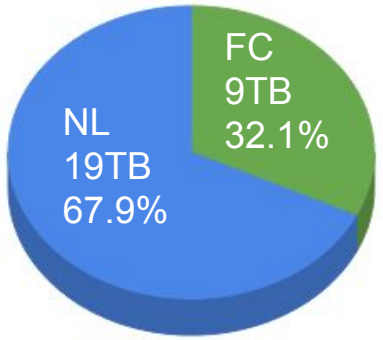
- Mountpoint no tiene que ser igual que el nombre
- Tamaño flexible (quota)

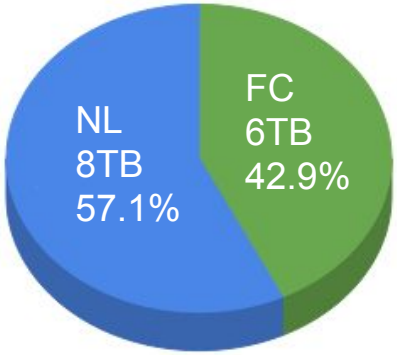
```
zfs set quota=300G pool1/nfs1_1
```

- Permisos de acceso directamente en ZFS

Distribución de pooles y recursos

	Tipo disco	*Tamaño*	*Usado*	*Nº vols*	*Nº recursos*	*Compresión*
SISTEMAS		**28TB**	**12.2TB**		**171**	
pool_prod_fc	Fast class	4TB	1.8TB	2x2TB	47	1.07%
pool_prod_nl	Near line	4TB	540GB	1x4TB	32	1.09%
pool_desa_fc	Fast class	2TB	186GB	1x2TB	34	1.22%
pool_desa_nl	Near line	2TB	503GB	1x2TB	51	1.26%
pool_sakai_fc	Fast class	3TB	2.37TB	2 (1+2TB)	3	1.11%
pool_bak_nl	Near line	12TB	6.5TB	3x4TB	3	off
pool_inter_nl	Near line	1TB	252GB	1x1TB	1	off
TELEMATICA		**14TB**	**6.12TB**		**35**	
pool_ryt_fc	Fast class	2TB	173MB	1x2TB	1	1.06%
pool_ryt_nl	Near line	8TB	4.15TB	2x4TB	27	1.00%
pool_box_fc	Fast class	4TB	1.97TB	2x2TB	7	1.14%
TV		**20TB**	**14.5TB**		**15**	
pool_tv_nl	Near line	20TB	14.5TB	5x4TB	15	off
SOPORTE		**2TB**	**985GB**		**2**	
pool_nov_nl	Near line	2TB	985GB	1x2TB	2	off





Propiedades de un recurso


```
# zfs get all pool_box_fc/umubox_dg_a1
NAME                        PROPERTY              VALUE                          SOURCE
pool_box_fc/umubox_dg_a1    type                  filesystem                     -
pool_box_fc/umubox_dg_a1    creation              mar nov 11 12:23 2014          -

pool_box_fc/umubox_dg_a1    mounted               yes                            -
pool_box_fc/umubox_dg_a1    mountpoint            /umubox_dg_a1                  local
pool_box_fc/umubox_dg_a1    sharenfs              rw=10.14.2.6:10.14.1.240/28    local
pool_box_fc/umubox_dg_a1    sharesmb              off                            default

pool_box_fc/umubox_dg_a1    quota                 400G                           local
pool_box_fc/umubox_dg_a1    used                  306G                           -
pool_box_fc/umubox_dg_a1    available             94,3G                          -

pool_box_fc/umubox_dg_a1    compression           lz4                            inherited from pool_box_fc
pool_box_fc/umubox_dg_a1    compressratio         1.11x                          -

pool_box_fc/umubox_dg_a1    atime                 off                            inherited from pool_box_fc

pool_box_fc/umubox_dg_a1    snapdir               hidden                         default

pool_box_fc/umubox_dg_a1    readonly              off                            default

pool_box_fc/umubox_dg_a1    dedup                 off                            default
```

Tareas habituales

Operaciones sobre pooles

- Crear un pool
- Ampliar pool existente
- Consultar estado de pooles y tareas en curso
- Exportar o importar (¡cuidado!)
- Eliminar un pool (¡cuidado!)

Operaciones sobre recursos

- Crear un recurso: nombre, tipo de disco (pool), tamaño, mountpoint, shares
- Modificar propiedades: shares, tamaño
- Modificar propiedades: mountpoint, nombre
- Eliminar un recurso (¡cuidado!)

Tareas habituales: Scrubs

- Chequeo de integridad del sistema de ficheros, recorre el contenido del pool en busca de errores
- Escanea los datos, no el total del pool
- Tarea en segundo plano, sólo una por nodo

```
# zpool status pool_name
  pool: pool_name
 state: ONLINE
  scan: scrub in progress since Fri Nov 11 12:52:35 2016
     1,22G scanned out of 40,9G at 47,9M/s, 0h14m to go
     0 repaired, 2,97% done
```

- Script genérico para realizar scrub de todos los pooles del filer
  - Planificado en CRON para un sábado del mes a las 0:00, según filer

Cuotas de usuario

- ZFS soporta cuotas de usuario
  - Establecimiento de cuotas y consulta mediante comandos locales
- No se integra con el comando “quota” para consultas remotas (via NFS)
- Creamos script para consulta y modificación de cuotas
  - Definir usuarios locales, certificados para autenticación SSH
  - Script en /home/usuario
  - Configurar /etc/sudoers para que puedan ejecutar comando /sbin/zfs

Failover cluster

- ZFS no ofrece mecanismo para clusterización
- Pooles montados exclusivamente en un nodo, pero visibles por todos
  - Usar “zpool import” y “zpool export” para montar los pooles en el nodo deseado
  - OJO: Cuidado con montar un pool en dos nodos a la vez, posible corrupción de datos
- Configuración de red equivalente en todos los nodos
- Scripts sencillos para failover
- El propio cliente NFS se recupera ante desconexiones puntuales

nas_telematica start

```
ifconfig eth2 10.59.0.4 netmask 255.255.252.0
ifconfig eth3 10.14.0.7 netmask 255.255.252.0 mtu 9000

zpool import pool_ryt_fc
zpool import pool_ryt_nl
zpool import pool_box_fc
```

nas_telematica stop

```
zpool export pool_ryt_fc
zpool export pool_ryt_nl
zpool export pool_box_fc

ifconfig eth2 down
ifconfig eth3 down
```

Snapshots y réplicas

- Gestión de snapshots integrada en el sistema de ficheros ZFS
  - Propiedad “snapdir” (hidden / visible) para acceder a los snapshots desde el recurso raíz
  - Snapdir: /fs_test/.zfs/snapshot/snap1
- Nombre de snapshot es igual al del recurso más un sufijo
- Propiedades del recurso snapshot reducidas
  - Propiedad “type” cambia de “filesystem” a “snapshot”
- Snapshots sucesivos siempre incrementales
- Eliminación de snapshots consolida para mantener la cadena
- Etiquetado de snapshots (hold, release)

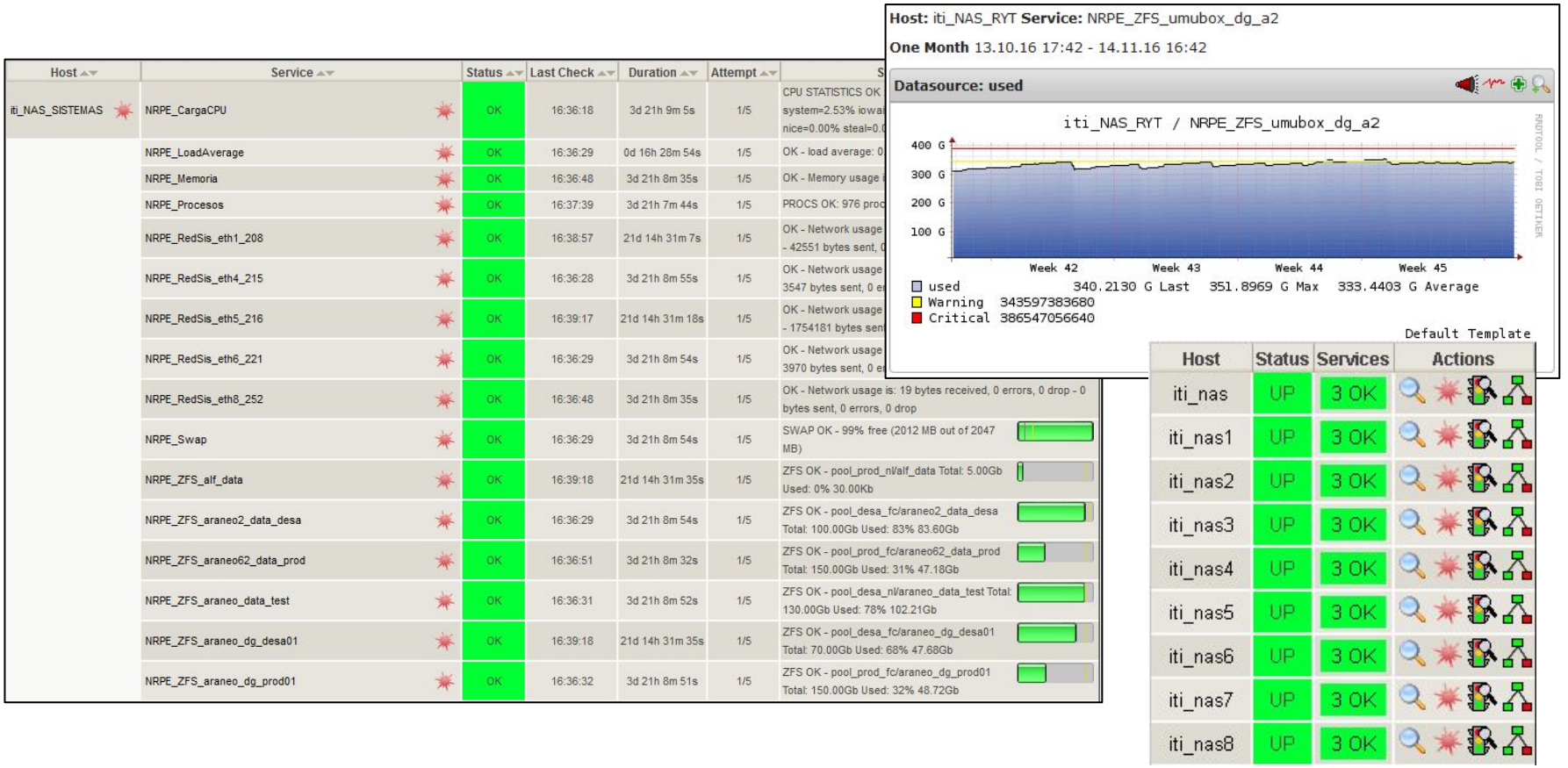
Snapshots y réplicas

- Mecanismo de réplicas basado en dos comandos
  - zfs send: convierte un recurso en un flujo de bytes
  - zfs receive: convierte un flujo de bytes en un recurso
- Los comandos se concatenan para realizar una réplica
- Cuando hay snapshots, se pueden hacer réplicas recursivas e incrementales
- Se puede encadenar con un SSH para enviar a un filer remoto

- Script para mantener estructura de snapshots y réplicas
  - Crear y eliminar snapshots, etiquetas de snapshot diario, semanal, etc
  - Réplicas con filer remoto (opcional)
  - Diferente retención en local y en remoto

Monitorización

- No hay check SNMP para ZFS, creamos scripts NRPE
- Monitorización de pooles, recursos, tráfico de red, procesos, memoria...



Host	Service	Status	Last Check	Duration	Attempt	S
iti_NAS_SISTEMAS	NRPE_CargaCPU	OK	16:36:18	3d 21h 9m 5s	1/5	CPU STATISTICS OK system=2.53% iowa nice=0.00% steal=0.0
	NRPE_LoadAverage	OK	16:36:29	0d 16h 28m 54s	1/5	OK - load average: 0
	NRPE_Memoria	OK	16:36:48	3d 21h 8m 35s	1/5	OK - Memory usage
	NRPE_Procesos	OK	16:37:39	3d 21h 7m 44s	1/5	PROCS OK: 976 proc
	NRPE_RedSis_eth1_208	OK	16:38:57	21d 14h 31m 7s	1/5	OK - Network usage - 42551 bytes sent, 0
	NRPE_RedSis_eth4_215	OK	16:36:28	3d 21h 8m 55s	1/5	OK - Network usage 3547 bytes sent, 0 e
	NRPE_RedSis_eth5_216	OK	16:39:17	21d 14h 31m 18s	1/5	OK - Network usage - 1754181 bytes sent
	NRPE_RedSis_eth6_221	OK	16:36:29	3d 21h 8m 54s	1/5	OK - Network usage 3970 bytes sent, 0 e
	NRPE_RedSis_eth8_252	OK	16:36:48	3d 21h 8m 35s	1/5	OK - Network usage is: 19 bytes received, 0 errors, 0 drop - 0 bytes sent, 0 errors, 0 drop
	NRPE_Swap	OK	16:36:29	3d 21h 8m 54s	1/5	SWAP OK - 99% free (2012 MB out of 2047 MB)
	NRPE_ZFS_alf_data	OK	16:39:18	21d 14h 31m 35s	1/5	ZFS OK - pool_prod_nl/alf_data Total: 5.00Gb Used: 0% 30.00Kb
	NRPE_ZFS_araneo2_data_desa	OK	16:36:29	3d 21h 8m 54s	1/5	ZFS OK - pool_desa_fc/araneo2_data_desa Total: 100.00Gb Used: 83% 83.60Gb
	NRPE_ZFS_araneo62_data_prod	OK	16:36:51	3d 21h 8m 32s	1/5	ZFS OK - pool_prod_fc/araneo62_data_prod Total: 150.00Gb Used: 31% 47.18Gb
	NRPE_ZFS_araneo_data_test	OK	16:36:31	3d 21h 8m 52s	1/5	ZFS OK - pool_desa_nl/araneo_data_test Total: 130.00Gb Used: 78% 102.21Gb
	NRPE_ZFS_araneo_dg_desa01	OK	16:39:18	21d 14h 31m 35s	1/5	ZFS OK - pool_desa_fc/araneo_dg_desa01 Total: 70.00Gb Used: 68% 47.68Gb
	NRPE_ZFS_araneo_dg_prod01	OK	16:36:32	3d 21h 8m 51s	1/5	ZFS OK - pool_prod_fc/araneo_dg_prod01 Total: 150.00Gb Used: 32% 48.72Gb

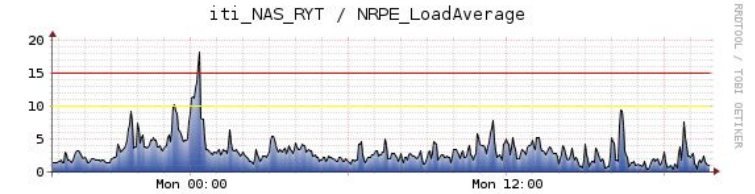
Host	Status	Services	Actions
iti_nas	UP	3 OK	
iti_nas1	UP	3 OK	
iti_nas2	UP	3 OK	
iti_nas3	UP	3 OK	
iti_nas4	UP	3 OK	
iti_nas5	UP	3 OK	
iti_nas6	UP	3 OK	
iti_nas7	UP	3 OK	
iti_nas8	UP	3 OK	

Monitorización

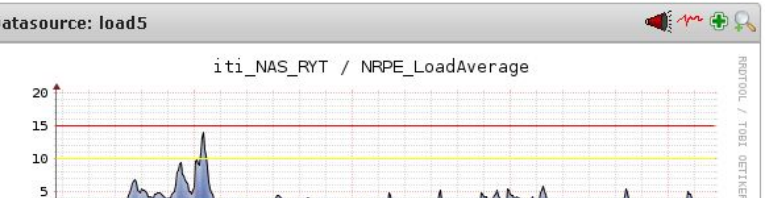
Host: iti_NAS_RYT **Service:** NRPE_LoadAverage

25 Hours 13.11.16 18:42 - 14.11.16 19:42

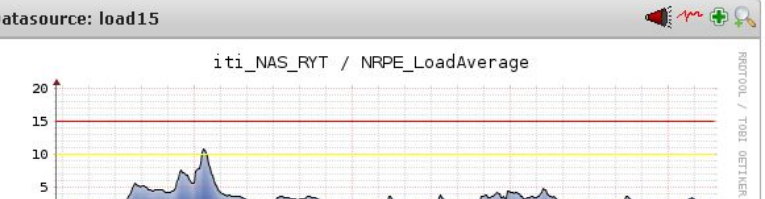
Datasource: load1



Datasource: load5



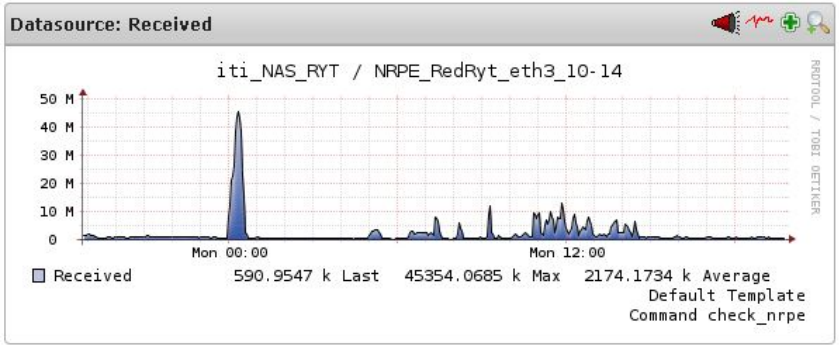
Datasource: load15



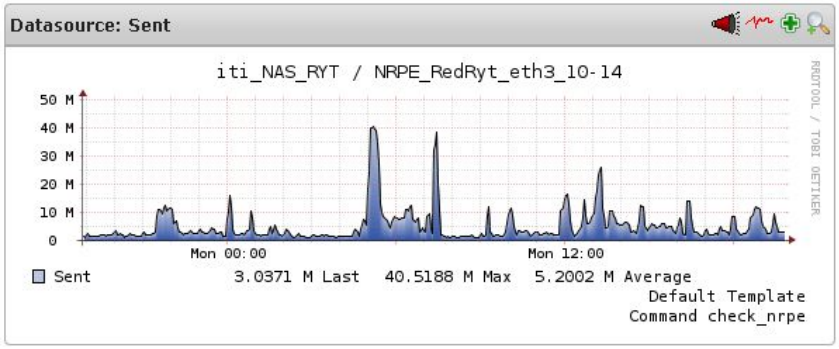
Host: iti_NAS_RYT **Service:** NRPE_RedRyt_eth3_10-14

25 Hours 13.11.16 18:47 - 14.11.16 19:47

Datasource: Received



Datasource: Sent

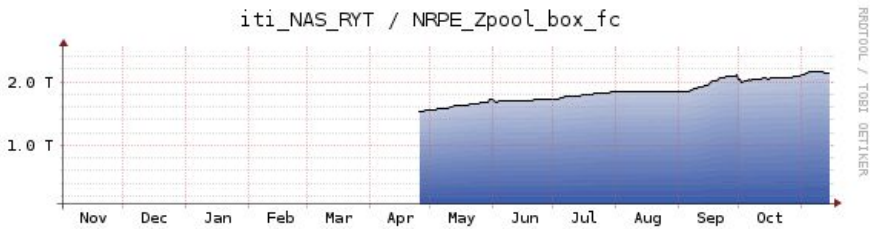


Monitorización

Host: iti_NAS_RYT **Service:** NRPE_Zpool_box_fc

One Year 31.10.15 20:01 - 14.11.16 20:01

Datasource: used

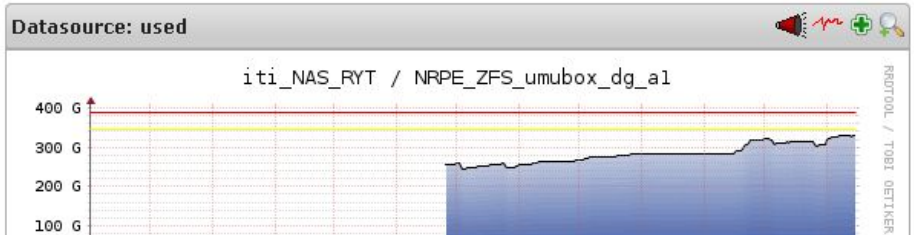


Datasource: size

Host: iti_NAS_RYT **Service:** NRPE_ZFS_umubox_dg_a1

One Year 31.10.15 20:09 - 14.11.16 20:09

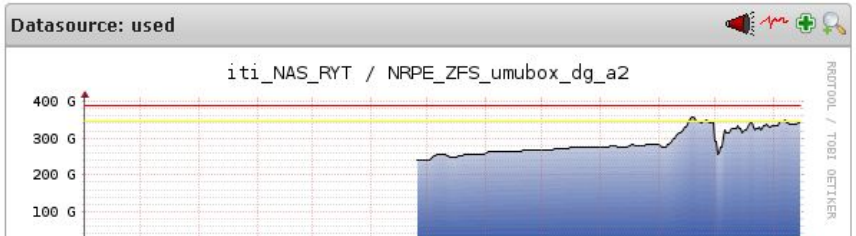
Datasource: used



Host: iti_NAS_RYT **Service:** NRPE_ZFS_umubox_dg_a2

One Year 31.10.15 20:11 - 14.11.16 20:11

Datasource: used



Servicios alrededor de NFS

- Araneo (web corporativa www.um.es)
- Configuraciones tomcat, OAS…
- Aula Virtual (SAKAI)
- DALI (impresión centralizada)
- Consigna (envío de archivos grandes por email)
- MySQL Telemática
- Socrates (disco compartido)
- Umubox (disco en la nube)
- Webs particulares departamentos y usuarios
- Concentrador de logs
- Diferentes entornos (producción y desarrollo)
- Archivo de videos y streaming en TV.UM

Tareas futuras

- Utilizar SSD para caché
- Snapshots y réplicas con CPD remoto
- Deduplicación
- Despliegue de nodos y configuración con Puppet
- Actualización de versiones de ZFS
- Automatización del cluster
- Posible entorno web para administrarlo

Gracias por su atención