Lustre

Álvaro Fernández / Alejandro Lamas / Javier Sánchez

Características

- Sistema de ficheros distribuido.
- Único espacio de nombres
  - /lustre/ific.uv.es/grid/atlas
- Escalable de Alto rendimiento:
  - Reparto de carga entre servidores
  - Reparto de ficheros entre OSTs.
- POSIX.
- Cuotas y ACLs

Características

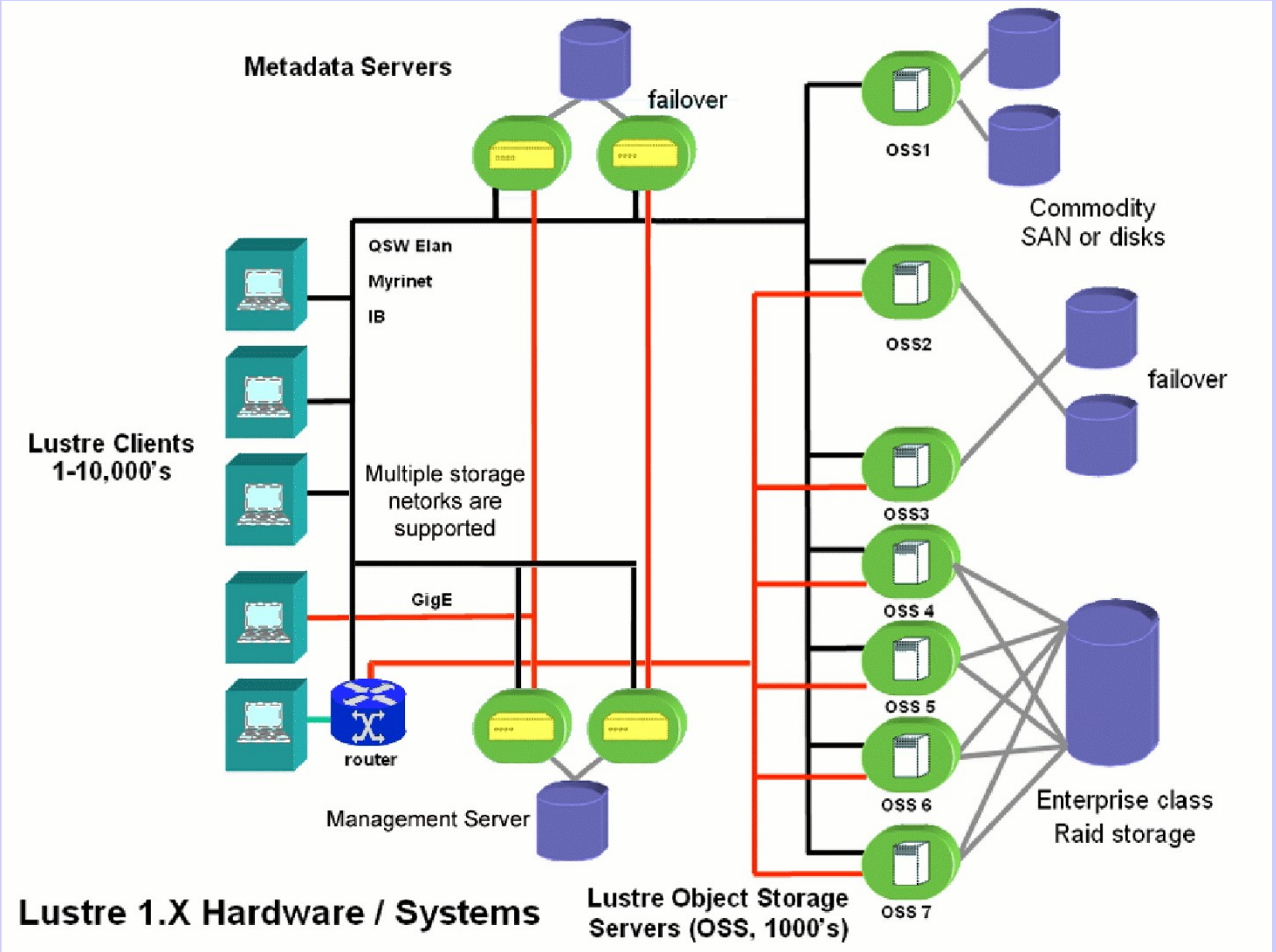
- Desarrollo activo (CFS->SUN->ORACLE):
  - ¿ORACLE? licencia GNU GPL.
  - Roadmap claro.
  - Mínimos retrasos en los plazos.
- Instalación sencilla:
  - Sin ficheros de configuración.
  - Disponibles módulos precompilados para clientes con kernel RHEL5.

Componentes básicos

- MGS: servidor de gestión.
  - Define la configuración de todos los sistemas de ficheros lustre de un cluster. Sólo habrá uno en el cluster.
- MDS: servidor de metadatos.
  - Un MDT por sistema de ficheros.
  - Recomendado: discos SAS (low seek time), raid 0, CPU 4 cores...
    - Journal externo.
- OSS: almacenamiento.
  - Exporta 1 ó más OST.
  - OSTs sobre discos, particiones, raids, volúmenes lógicos... (8TB max)
- Clientes.
  - Acceden al sistema de ficheros mediante kernel específico o usando módulos precompilados (RHEL)

Se puede hacer cualquier combinación de componentes (pruebas).

Componentes básicos



Configuración actual

- 1 metadirectorio
- 36 OSTs en 6 servidores (OSS)
- 66 clientes

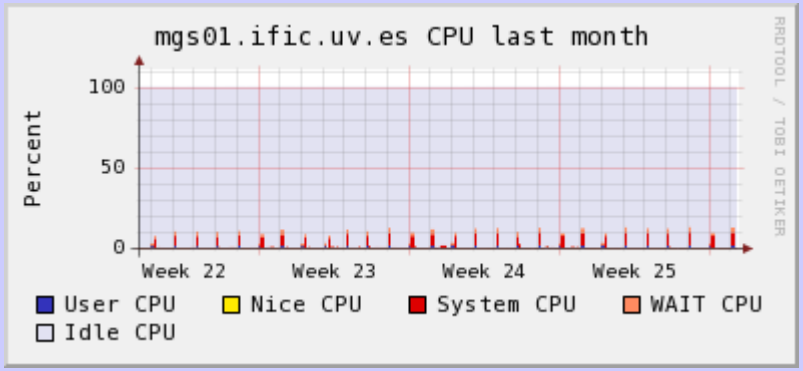
MGS+MDT

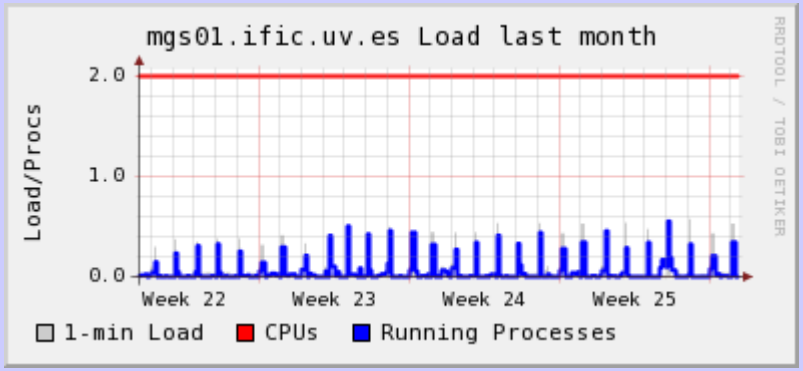
- Pentium D 3.2 GHz
- 4 GB RAM
- HD RAID-1 por hardware

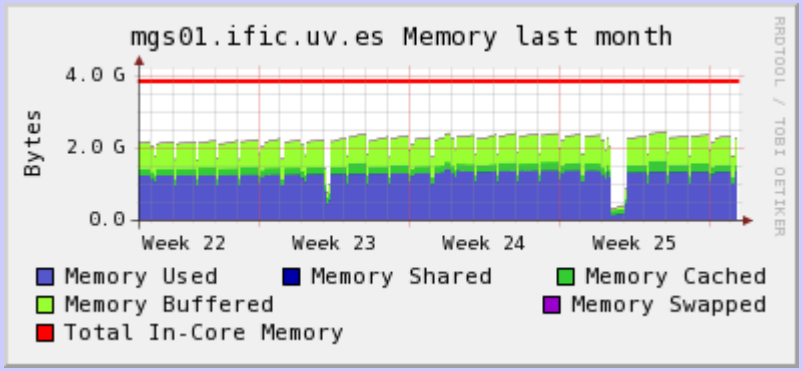
- No esta muy cargado y la respuesta es buena
- Importante I/O y número de conexiones tcp por servidores y clientes.

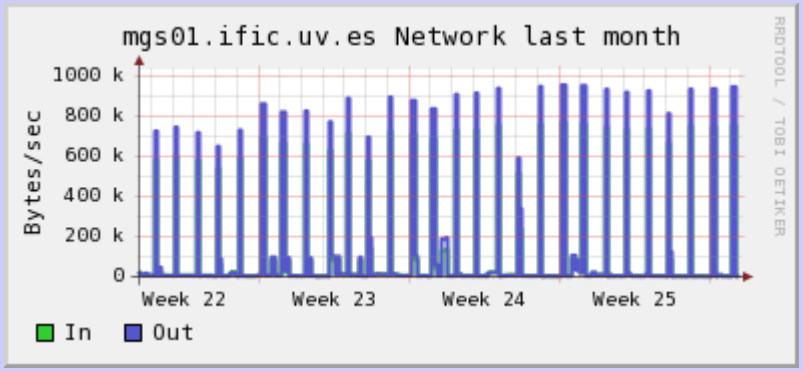
- Migración a un sistema HA.

MDT+MGS status









Instalación

- Software:
  - Software lustre:
    - kernel:
      - kernel-lustre-smp-2.6.9-42.0.10.EL_lustre_1.6.0.1.x86_64.rpm
    - módulos del kernel:
      - lustre-modules-1.6.0.1-2.6.9_42.0.10.EL_lustre_1.6.0.1smp.x86_64.rpm
    - utilidades lustre:
      - lustre-1.6.0.1-2.6.9_42.0.10.EL_lustre_1.6.0.1smp.x86_64.rpm

- Actualización trivial.
  - Instalar nuevas versiones y reiniciar.
  - Compatibilidad hacia atrás entre versiones.

Instalación

- Software:
  - e2fsprogs-1.39.cfs2-0.x86_64.rpm (soporte EA)
  - dump-0.4b41-1.x86_64.rpm
  - rmt-0.4b41-1.x86_64.rpm
  - acl-2.2.39-1.1.x86_64.rpm
  - libacl-2.2.39-1.1.x86_64.rpm
  - attr-2.4.28-1.2.x86_64.rpm
  - libattr-2.4.28-1.2.x86_64.rpm

No necesario en los clientes.

Configuración

- Deshabilitado selinux en todos los nodos.
- Todos los nodos han de estar sincronizados.
  - NTP
- UID/GID global en el cluster.
  - En caso de tener grupos secundarios:
    - /proc/fs/lustre/mds/mds-service/group_upcall
    - Herramienta básica que lee passwd & group en MDS
      - Necesario tener los usuarios y grupos en el metadirectorio.
    - Información en caché 5 min (ajustable).
- Cortafuegos.

/proc/fs/lustre

Posibilidad de ajustar:

- Timeouts.
- Read-ahead (por cliente).
- Cantidad de datos en el caché (por cliente).
- etc.

Listas de control de acceso

- POSIX.
- Se configuran en la creación del MDT o en su montaje (EA).
- Los clientes no necesitan configuración.
- `setfacl, getfacl y chacl`.

- Usadas para implementar las políticas de acceso de ATLAS.

Listas de control de acceso

- Esquema permisos ATLAS.

	mc	user	
	r:atlas w:prod	r:atlas w:atlas	
	atlp000	atlu000	← mapping
owner →	storm g:atlp:rwx g:atlas:r-x	storm g:atlp:rwx g:atlas:rwx	← acls ATLAS
	g:atlp	g:atlas	← mapping
	g:atlp:rwx g:atlas:r-x	g:atlas:rwx	← acls LUSTRE ← default acl

Cuotas

- Se administran con el comando lfs.
- No es necesario usar parámetros al montar el sistema de ficheros.
- Se pueden establecer para usuarios/grupos.

- ¡El Tier-3 necesita cuotas!
  - Los usuarios usan todo lo disponible sin preguntar.

Striping

- Ventajas: mayor ancho de banda y creación de ficheros de mayor tamaño que la capacidad de almacenamiento de un sólo OST.
- Puede configurarse para directorios y para ficheros individuales.
- No lo usamos.
- Provoca una mayor carga sobre el sistema y el riesgo de perder una gran cantidad de ficheros al perder un OST no compensa.

Múltiples sistemas de ficheros.

- Se pueden montar varios sistemas de ficheros lustre en un mismo cliente, unos sobre otros.

- Ejemplo:
  - `mount -t lustre wn173@tcp0:/ificfs /lustre/ific.uv.es/`
  - `mount -t lustre wn173@tcp0:/tier2 /lustre/ific.uv.es/tier2/`
  - `mount -t lustre wn173@tcp0:/tier3 /lustre/ific.uv.es/tier3/`

Backup

- Usando dump.
- Copia en disco.
- Sólo hacemos backup de MGS/MDS.
  - MGS: 41 KB en 1 segundo.
  - MDT: 121 MB en 4 minutos.

Acceso almacenamiento

- StoRM + GridFTP
  - Simples clientes lustre (rw).
  - Mapeo de usuarios al usuario 000 del pool.
    - atlu000 todos los usuarios de ATLAS.
    - atlp000 todos los usuarios de producción.
    - atls000 gestión software.
  - No comparten el vomsdir.

  - Protocolos:
    - gridftp
    - file (WNs,UIs)
- Otros servicios.
  - Acceso web mediante certificado grid.
  - http://seview.ific.uv.es

Acceso almacenamiento

- GridFTP
  - EGEE no soporta un SE clásico
  - Instalacion con YAIM 'parcheada ligeramente'
  - Servidor Gridftp de VDT GT4 (pre web services)
  - vdt_globus_data_server-VDT1.6.1x86_rhas_4-7.i386.rpm

  - Cualquier perfil que configure un gridftp (+VOMS, etc) es valido para proporcionar el acceso a Lustre.

Conclusiones

- POSIX.
- Administración sencilla.
- Facilidad para añadir nuevos nodos en caliente.
- Integración grid sin excesivas complicaciones.
  - gridftp sencillo.
  - StoRM menos sencillo.

Configuración de los nodos

- Nodo central
  - MGS:
    - `mkfs.lustre --mgs /dev/sda5`
    - `mkdir -p /mnt/mgs`
    - `mount -t lustre /dev/sda5 /mnt/mgs`
  - MDT:
    - `mkfs.lustre --fsname=ificfs --mdt --mgsnode=wn173@tcp0 --mountfsoptions=acl /dev/sda6`
    - `mkdir -p /mnt/ific/mdt`
    - `mount -t lustre /dev/sda6 /mnt/ific/mdt`
  - Entradas en el fichero /etc/fstab.
    - `LABEL=MGS  /mnt/mgs  lustre  defaults,_netdev 0 0`
    - `LABEL=ificfs-MDT0000  /mnt/ific/mdt  lustre  defaults,_netdev 0 0`

Configuración de los nodos

- Servidores de almacenamiento (OSS)
  - OSTs (añadibles en caliente).
    - `mkfs.lustre --fsname=ificfs --ost --mgsnode=wn173@tcp0 /dev/sda5`
    - `mkdir -p /mnt/ific/ost0`
    - `mount -t lustre /dev/sda5 /mnt/ific/ost0`
  - Entradas en el fichero /etc/fstab:
    - `LABEL=ificfs-OST0000  /mnt/ific/ost0  lustre  defaults,_netdev 0 0`
  - SUN x4500 (6 OSTs)

- Clientes
  - `mkdir -p /lustre/ific.uv.es`
  - `mount -t lustre wn173@tcp0:/ificfs /lustre/ific.uv.es`
  - Entradas en el fichero /etc/fstab.
    - `wn173@tcp:/ificfs  /lustre/ific.uv.es/  lustre  defaults,_netdev 0 0`
  - SRM, GridFTPs, WNs, UIs...

Operaciones con OST's

- Fallo de un OST.
  - Parar OST (permanente).
    - `lctl conf_param testfs-OST0001.osc.active=0`
    - Provoca error de entrada/salida inmediato en lectura. Las operaciones de borrado se guardan y realizan nada más activarse. Las escrituras irán a parar a los otros OST
  - Reactivar OST.
    - `lctl conf_param testfs-OST0001.osc.active=1`
    - `/proc/fs/lustre/obdfilter/ificfs-OST0001/recovery_status`
  - Desactivar (temporal).
    - `lctl --device 5 deactivate`
    - `Evita la creación de ficheros, pero permite la lectura.`
  - Activar
    - `lctl --device 5 activate`

Cuotas

- Permiten la administración del espacio para un experimento o usuario.
- Ejemplos de uso:
- `lfs quotacheck  /lustre/ific.uv.es/`
- `lfs setquota -g dteam 100000 200000 0 0 /lustre/ific.uv.es/`
- `lfs setquota -g dteam 0 0 0 0 /lustre/ific.uv.es/`
- `lfs quota -g dteam /lustre/ific.uv.es/`

Striping

- Ejemplos de uso:
  - `lfs setstripe /lustre/ific.uv.es/grid/dteam/ <stripe-size> <ost_index-start> <ost_number-to-use>`
  - `lfs setstripe /lustre/ific.uv.es/grid/dteam/ 0 -1 0`
  - `lfs getstripe /lustre/ific.uv.es/grid/dteam/`