Universidad Nacional Autónoma de México.

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Programa de Posgrado en Ciencia e Ingeniería de la Computación

Sistema de almacenamiento para el laboratorio nacional de observación de la tierra

TESIS

que para obtener el grado de

Especialista en Computo de Alto Rendimiento

PRESENTA

Alejandro Aguilar Sierra

TUTOR PRINCIPAL DE TESIS

Lukas Nellen Filla

Ciudad Universitaria, 2020

Sistema de almacenamiento para el Laboratorio Nacional de Observación de la Tierra

Alejandro Aguilar Sierra

7 de agosto de 2020

QUE PARA OBTENER EL GRADO DE:

DIRECTOR:

Resumen

En el Laboratorio Nacional de Observación de la Tierra (LANOT) se reciben diariamente varios terabytes de información de satélite y se genera una cantidad similar de datos en productos derivados. Se requiere un sistema para almacenar esa información por un tiempo determinado y en casos especiales, ser archivada por tiempo indefinido, además de organizar cantidades masivas de información de manera eficiente y fácil de localizar para el usuario final.

En este trabajo se reporta la implementación de un sistema de bajo costo y alto rendimiento para resolver este problema, con uso preferencial de sistemas de código abierto y el aprovechamiento de años de experiencia acumulada en desarrollo de soluciones a problemas similares en otras dependencias de la UNAM, especialmente en el Instituto de Ciencias Nucleares.

Índice general

Capítulo 1

Introducción

1.1. Problema

El Laboratorio Nacional de Observación de la Tierra, LANOT [Agu18], es un laboratorio nacional CONACyT creado con el apoyo de un consorcio formado por instituciones tanto dentro como fuera de la UNAM. Tiene tres sistemas de adquisión satelital: GOES 16 Rebroadcast [NN17], GeonetCast [Ear] y Polar (satélites SNPP, JSPP, Aqua, Terra y EUMETSAT MetOp). Los datos recibidos por esos sistemas, son sometidos a distintos niveles de procesamiento y se ponen a disposición de los usuarios del LANOT, entre ellos dependencias federales como el CENAPRED, la SEMAR, el Servicio Meteorológico Nacional y educativas como la Universidad Autónoma del Estado de México o la misma UNAM. En el LANOT se reciben diariamente más de 2 terabytes de información de satélite y se genera una cantidad similar de datos en productos derivados. Los sistemas que se tenían antes de este proyecto, tenían capacidad de almacenamiento limitada por lo que unicamente se guardaba la información las primeras 12 horas y después se eliminaba, a menos que se hubiera respaldado manualmente en un disco externo. Si bien en la mayoría de los casos las imágenes de satélite se almacenan en algún centro de datos en el país de origen del satélite correspondiente, es conveniente un sistema para recuperar datos de manera oportuna, sobre todo para casos de alguna contingencia ambiental del tipo incendio forestal, erupción volcánica, tormenta meteorológica o invasión de sargazo en las playas.

Por otro lado, la mayor parte del procesamiento de dichos datos se realiza con software propietario que no solamente depende de una costosa licencia, sino que los algoritmos y los programas que se utilizan están en código ejecutable cerrado que no se puede modificar. Además, se ha observado que la georreferenciación de los productos finales tiene una precisión apropiada para interpretación visual pero no para utilizar los datos con alto nivel de detalle y precisión geográfica. Sin el acceso directo a los algoritmos, no es posible controlar la calidad de los productos generados.

En este trabajo se propone un sistema de bajo costo y alto rendimiento

para resolver por un lado el problema del almacenamiento y por el otro, el procesamiento de productos derivados con software de código abierto.

1.2. Propuesta de Solución

Ante el problema descrito en la sección anterior, el laboratorio solicitó y recibió cotizaciones de soluciones comerciales por algunos proveedores, por montos promedio del orden de 5 millones de pesos por doscientos terabytes, es decir, del orden de 25,000 pesos por terabyte.

Afortunadamente, otros institutos de la UNAM se han enfrentado con problemas similares de almacenamiento masivo y han encontrado soluciones alternativas. En particular, en la Unidad de Cómputo del Instituto de Ciencias Nucleares se cuenta con experiencia de más de 10 años en la construcción y administración de sistemas de almacenamiento de alto rendimiento [MP17]. Su sistema actual es parte de la infraestructura de cómputo del Laboratorio Nacional HAWC (High Altitude Water Cherenkov Gamma Ray Observatory) [P H19; A S15; A U12] cuyo responsable técnico es el Dr. Lukas Nellen. Durante el curso «Laboratorio de Clusters y GRIDS» que llevé con los profesores Lukas Nellen, Luciano Diaz y Eduardo Murrieta, tuve oportunidad de conocer de primera mano su sistema y durante el mismo curso desarrollamos una versión virtual de un sistema de almacenamiento.

Con base en dicha experiencia y la valiosa asesoría de mis profesores, diseñé una propuesta de sistema de almacenamiento paralelo de bajo costo y alto rendimiento, estable y altamente escalable, basado en software de código abierto (sistema operativo Linux y sistema de archivos distribuido Lustre) y hardware armado, sumado a la infraestructura que ya existe en el laboratorio, en red local de alta velocidad (10Gb).

Después de solicitar cotizaciones con diferentes proveedores de la UNAM, estimamos que esta solución ad-hoc con equipo armado bajo diseño y software libre, tendría un costo aproximado de un millón de pesos por 500 terabytes, es decir, 2,000.00 pesos por terabyte, lo cual representa solo un 8 % del costo comercial promedio.

Es necesario remarcar que el software libre no es gratuito, lleva un costo oculto en mayor responsabilidad del usuario y personal mejor capacitado para construir y operar el sistema, pero esa es precisamente nuestra aportación.

1.3. Sistema de Almacenamiento LUSTRE

Lustre[Bra18] es un sistema de archivos distribuido de alto rendimiento y es utilizado por la mayoría de los sistemas de supercómputo en el mundo por su gran rendimiento y escalabilidad [MP17]. Es un sistema distribuido que requiere la participación de varios servidores conectados en una red local de alta velocidad. Aunque es un conjunto de servidores, un cliente conectado al sistema lo verá como un disco con un solo sistema de archivos.

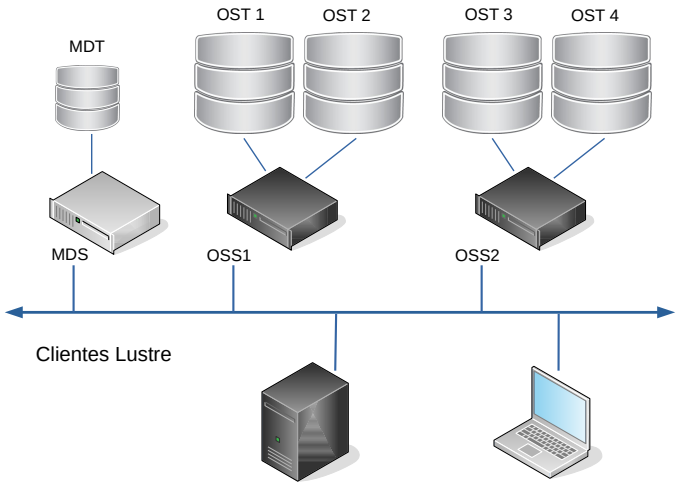

Figura 1.1: Arquitectura del sistema de archivos Lustre

Lustre se compone de tres tipos de servidores, los clientes, capaces de escribir y recuperar archivos, los responsables de asignar espacio de almacenamiento y organizar el espacio de nombres de los archivos (servidores de metadatos) y los responsables de la información en sí (servidores de almacenamiento de objetos), que se dividen en los siguientes componentes:

Object Storage Target (**OST**) Medio físico de almacenamiento de datos, que es un disco o un conjunto de discos en RAID (Redundant Array of Independent Disks). Un RAID permite organizar un conjunto de discos como una unidad lógica y manejar distintos tipos de redundancia para minimizar fallas y pérdida de datos.

Servidor de Objetos de Almacenamiento (**OSS**) Suministra almacenamiento en masa del contenido de los archivos del sistema. Los datos se almacenan físicamente en los OST y un solo OSS puede manejar varios OSTs.

Metadata Target (**MDT**) Dispositivo de almacenamiento para la información de metadatos.

Servidor de Metadatos (**MDS**) Proporciona servicios de metadatos y el espacio de nombres del sistema. Almacena la jerarquía de directorios y la información de los archivos, como las fechas de acceso, el tamaño, etc. Si este servidor falla, no será posible recuperar un archivo en el depósito.

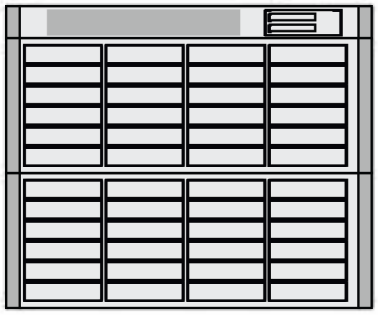

Figura 1.2: Representación en diagrama del OSS (cajón delgado superior) y sus OSTs repartidas en dos unidades de almacenamiento (gabinetes con 24 discos cada uno), controladas por la tarjeta RAID del OSS.

Un solo sistema Lustre puede escalar a cientos de OSSs, cada uno con varios OSTs y controlados por un MDS.

1.3.1. Unidad de almacenamiento físico

En los centros de datos se acostumbra colocar equipo de cómputo en un bastidor metálico de medidas estandarizadas, conocido por su nombre en inglés, *rack*. La altura vertical de los equipos que pueden colocarse en un rack se mide en unidades de rack o simplemente U y equivale a 1.75 pulgadas o 4.445 cm.

En el sistema que proponemos, cada unidad de almacenamiento (compuesta por uno o varios OSTs) constará de un chasis de 4U con espacio para 24 discos de 12 TB c/u, 288 TB en total, aunque al organizarse físicamente en un arreglo RAID 6, con dos discos de paridad y uno de *spare* (repuesto), esa capacidad se reduce a poco menos que 250 TB. Estas unidades se denominan JBOD (*just a bunch of disks*) y no requieren una tarjeta madre puesto que son controlados por una tarjeta RAID conectada a su panel trasero.

El OSS que maneje una o más unidades de almacenamiento, requiere contar con una tarjeta madre con un buen procesador, al menos 64 GB de memoria RAM y una tarjeta controladora RAID con la que controla las unidades de almacenamiento que tenga conectadas en cascada a la tarjeta RAID. Si el OSS se instala en el mismo gabinete de una unidad de almacenamiento, la convierte en una «unidad de almacenamiento inteligente».

1.3.2. Escalabilidad

Para aumentar la capacidad de almacenamiento del sistema, simplemente se agregan más JBODs como la descrita en la sección anterior y si hiciera falta, sus correspondientes OSS. El MDS las reconocerá y agregará al sistema de manera incremental.

1.3.3. Prevención de fallas

Siempre es conveniente prevenir las fallas que podrían comprometer la integridad del sistema de almacenamiento, a distintos niveles, desde fallas externas, como una interrupción eléctrica, a internas, como discos dañados.

1.3.3.1. Falla física de un disco

Cada unidad física contará con arreglo RAID 6, que implica reservar dos discos de paridad, y un disco de reserva, o *spare*. Tendrían que fallar a la vez 2 discos para que se comprometa el sistema. Cuando falle un disco, será reemplazado de inmediato por el de reserva, sin perturbar el sistema. Ese disco debe ser repuesto lo antes posible, para que siempre haya un disco de reserva en el sistema.

1.3.3.2. Robustez de la red interna

El cluster de almacenamiento contará con su propio switch de 10Gb lo que permitirá altas velocidades dentro del cluster y no dependerá de una red externa para funcionar. El switch a su vez podrá ser monitoreado desde el servidor principal (ver sección 2.4).

1.3.3.3. Falla eléctrica

El sistema se colocará en un rack con UPS (*interruptible power supply*, protector eléctrico con batería) redundante y los servidores y JBODs cuentan con fuente de poder redundante. Esto prevendrá problemas debidos a inestabilidad de la red eléctrica y en el caso de falla de una de las fuentes de alimentación de los equipos. Además, se cuenta con el sistema de respaldo eléctrico del Instituto de Geografía, consistente en un generador diesel.

1.3.3.4. Detección temprana

Se considera incluir mecanismos que nos permitan monitorear el funcionamiento del sistema de manera remota y automática. A nivel hardware, lo más básico es prevenir fallas eléctricas, como apagones y saltos de voltaje, lo que se logra con la UPS, que se comunica con el servidor principal. En caso de agotamiento de la batería, se procederá a apagar los demás servidores del clúster de manera controlada.

También se prevee que los discos duros en algún momento empezarán a fallar y convendrá contar con un sistema que pueda identificar rápidamente si algún disco está fallando, para ser reemplazado. Esto se puede hacer con el sistema de tecnologías de monitoreo, análisis y reporte (S.M.A.R.T.) de los dispositivos de almacenamiento y las tarjetas controladoras.

A nivel sistema operativo se cuentan con herramientas para monitorear el correcto funcionamiento de los procesos de este nivel, como que los módulos

del kernel y la versión del kernel correctos hayan sido cargados al momento de arranque del sistema.

Finalmente, para monitorear los procesos de operación, se usará correo electrónico local y herramientas comunes de monitoreo de clústers.

1.3.4. Seguridad

El sistema estará accesible únicamente para servidores dentro de la red local, conectados por el switch en una red privada. Los clientes accesibles desde el exterior y que tengan acceso al sistema, tendrán su propio *firewall* para evitar intrusiones. En los servidores con sistema operativo Linux CentOS 7, se usará `iptables` y en sistemas con GNU/Linux Debian, el más moderno `nftables`.

1.4. Red local completa

Una vez descritos los elementos, la configuración de la red local de alta velocidad propuesta, se ilustra en el siguiente diagrama (figura 1.3). En esta red, los nodos de procesamiento tendran acceso directo al sistema de almacenamiento por medio de una red de alta velocidad de 10 Gbs.

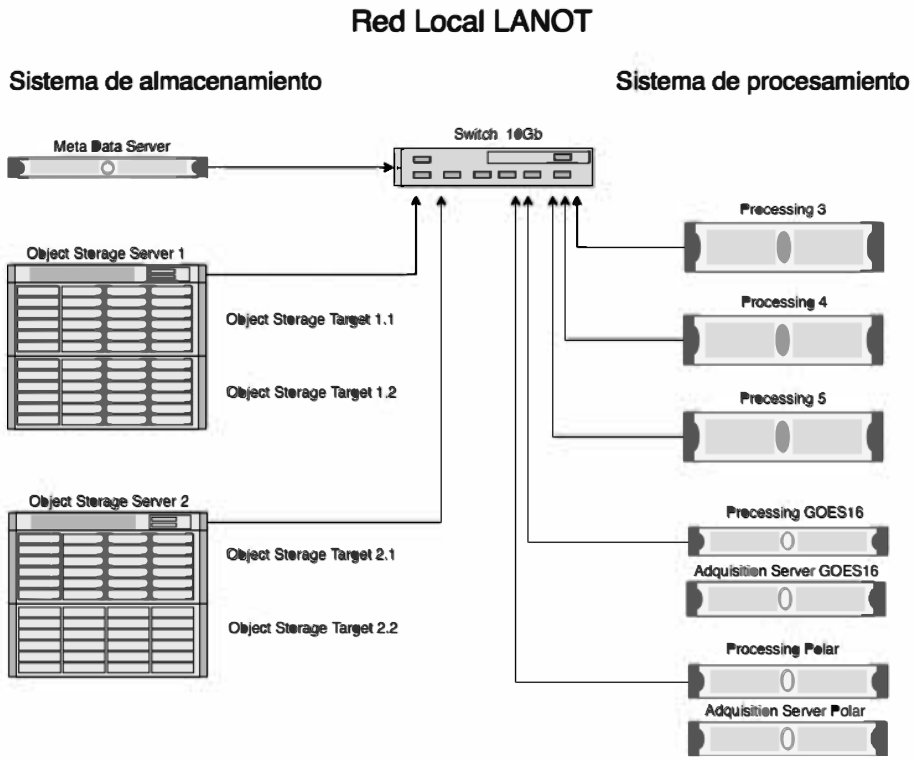

Figura 1.3: Diagrama de la red local propuesta

Capítulo 2

Implementación

Después de considerar distintas cotizaciones de proveedores autorizados por la UNAM y recomendados por nuestros asesores, se eligió adquirir equipo armable bajo diseño de la marca SuperMicro. En este capítulo usaremos la nomenclatura descrita en el capítulo anterior para explicar cómo se configuró finalmente el sistema de almacenamiento con base en Lustre.

2.1. Hardware

Para contener el sistema, se adquirió un rack de 42 unidades rack (42U) con una altura de 1.99m y para redundancia y protección eléctrica, una UPS de 15 KVA y una UPS de 5 KVA. Los chasis de los servidores y JBODs contienen cada uno fuente de poder redundante (dos fuentes), cada una conectada a una de las dos UPS, de modo que si una falla, se siga usando la otra.

De acuerdo a la propuesta del capítulo anterior, se adquirieron cuatro cajones (chasis) 4U con cupo para 24 discos. Los cajones pueden ser «inteligentes», es decir, incluir una tarjeta madre con procesador y memoria dentro del mismo cajón, o «esclavos», con solamente una tarjeta *backplane* controlada por una tarjeta RAID externa.

En la figura 2.1 se muestra una fotografía del rack con el equipo instalado.

2.1.1. OSS

La tarjeta madre del OSS contiene

- Procesador Intel Xeon Silver 4114, 2.20GHz. Se eligió este procesador por su robustez y porque brinda el poder de procesamiento adecuado.
- Memoria RAM 64GB, DDR4, 2666 MT/s, EEC, más que suficiente para operar el sistema.

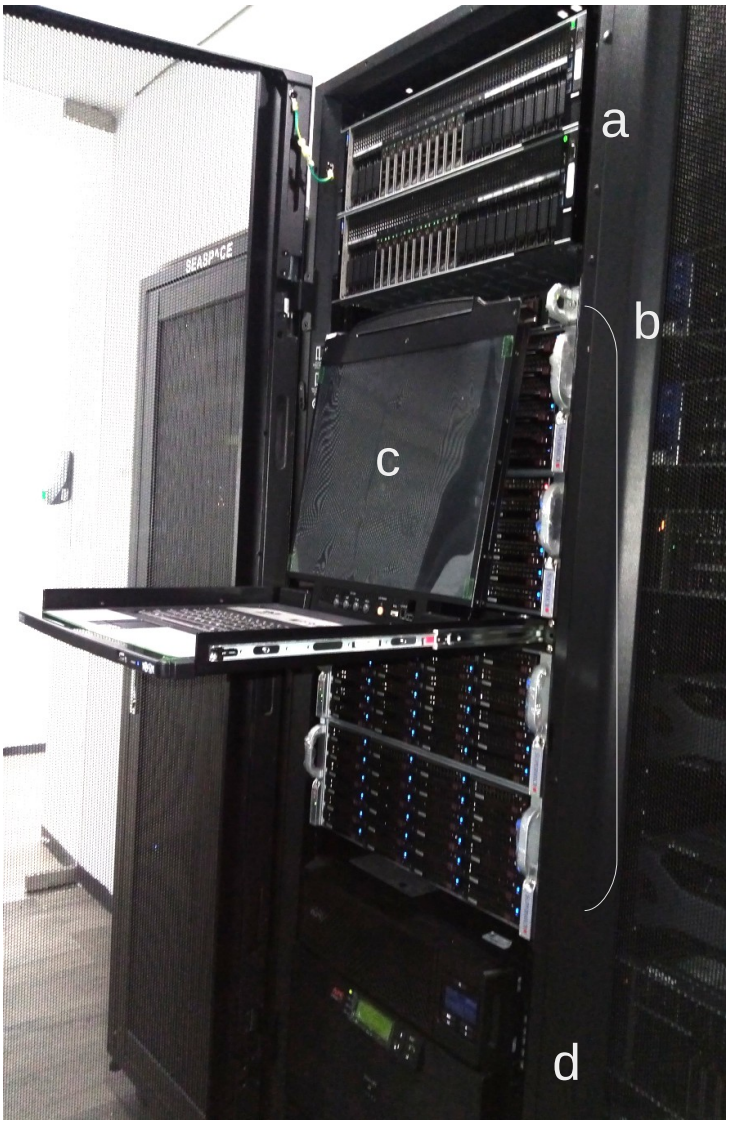

Figura 2.1: Rack con los servidores de procesamiento (a), el sistema de almacenamiento (b), una consola de administración de todos los equipos (c), y los UPS (d).

- Disco NVMe 128GB de estado sólido para el sistema operativo. Se decidió usar esta tecnología por su rapidez y ausencia de piezas móviles y porque si llegara a fallar, puede reemplazarse fácilmente sin comprometer la informacion en los discos mecánicos, que se utilizarán únicamente para el sistema de almacenamiento.
- Tarjeta controladora RAID, modelo LSI MegaRAID SAS 9361-8i SATA/SAS High Performance 12Gb/s, con batería, que permitirá organizar los discos en los OST como unidades virtuales.
- Tarjeta de red modelo Intel Corporation Ethernet Connection X722, dos puertos, con velocidad de 10GB.

Dicha tarjeta madre reside en un cajón 4U que contiene OSTs, lo que lo hace el cajón «inteligente», de acuerdo a la definición de la sección 1.3.

2.1.2. OSTs

Los OSTs se armaron como se describe en la sección 1.3.1:

- 24 discos de 12TB, 7200 rpm, 256MB Cache SAS-III, 12Gb/s, modelo 3.5" Seagate HDD ST12000NM0027.
- Un gabinete para los 24 discos, para una capacidad cruda aproximada de 250TB.
- En cada cajón se configuró un RAID 6 + spare, en el que 2 de los 24 discos se asignan para paridad y uno como repuesto, de manera que si un disco resulta dañado, el repuesto entrará en operación.
- Por las limitaciones de un sistema de archivos Lustre (basado en una extensión del sistema de archivos más común en Linux, `ext4`), es necesario agrupar los discos en unidades lógicas de 50TB, por medio de la tarjeta controladora RAID. Así los discos físicos se agruparon en conjuntos de 4 discos y fracción, para formar un volumen lógico de 50TB, para un total de 5 de estos volúmenes lógicos por gabinete y por lo tanto, 5 OSTs. Esta organización es a un nivel superior a la configuración de RAID 6 del inciso anterior y no le afecta.

Como ya se mencionó, si el cajón incluye una tarjeta madre, ahí se alojará el OSS y lo hará un cajón «inteligente». Un cajón esclavo (sin tarjeta madre) es controlado por la controladora RAID que reside en el cajón inteligente. En nuestro diseño, combinamos un cajón inteligente con uno esclavo, con lo que se tiene un total de 10 OSTs virtuales, con lo que se alcanza una capacidad aproximada de 500TB o medio Petabyte por conjunto.

En esta implementación adquirimos y armamos dos conjuntos OSS con un total cercano a 1 PB de capacidad de almacenamiento.

Figura 2.2: El MDS (el servidor 1U de arriba) y un par inteligente y esclavo (JBOD), el OSS alojado en el gabinete 4U de enmedio y los OSTs distribuidos en ambos gabinetes 4U. El foco rojo indica el disco de repuesto en cada gabinete.

2.1.3. MDS

Para el MDS no se requiere mucho espacio, por lo que se instaló en un cajón delgado 1U con elementos similares al OSS, con el doble de memoria RAM:

- Procesador Intel Xeon Silver 4114, 2.20GHz
- 132GB de memoria RAM, DDR4, 2666 MT/s, EEC.
- Disco NVMe 256GB para el sistema operativo.
- 5 discos marca Kingston, modelo SUV500480G, SSD SATA III, 2.5", 480GB, organizados 4 en RAID 10 y uno de repuesto, para mayor seguridad y redundancia, pues ahí se almacenarán los metadatos del sistema de almacenamiento y si llegara a fallar, se perdería el acceso a los datos almacenados.
- Tarjeta controladora RAID, modelo LSI MegaRAID SAS 9361-8i SATA/SAS High Performance 12Gb/s, con batería.
- Tarjeta de red modelo Intel Corporation Ethernet Connection X722, dos puertos, con velocidad de 10GB.

2.2. Software

2.2.1. Sistema operativo

En los servidores de almacenamiento descritos en la sección anterior, se instaló el sistema operativo Linux con la distribución CentOS 7 y los paquetes básicos para un sistema en red. También se instalaron los paquetes necesarios para echar a andar un sistema Lustre, como herramientas para el manejo de sistemas de archivos, módulos del kernel, etc.

Los clientes pueden tener otro sistema operativo, siempre y cuando cuenten con las bibliotecas y módulos del kernel de clientes de Lustre, es decir que puedan establecer acceso a un sistema distribuido de archivos Lustre.

El cliente principal es el nodo maestro de procesamiento, llamado `cirrus`, al que también se le instaló el sistema operativo CentOS7 y que copiará los archivos que vayan llegando a las estaciones de recepción, los procesará y los resultados los almacenará. En un servidor secundario llamado `stratus` se instaló el sistema Linux Debian 9.9 con el kernel 4.9 y se compilaron los módulos Lustre del kernel, que aunque no son las versiones más recientes, son compatibles con la versión de Lustre que estamos usando.

Ver la lista detallada en el apéndice B.

2.3. Configuración

2.3.1. Servidores de almacenamiento

Se siguieron los procedimientos correspondientes para configurar el MDS y los OSS de Lustre, para brindar el servicio de almacenamiento distribuido. En nuestro caso fueron fundamentales las notas del curso *Clusters y Grids* de la especialidad y la asesoría de mis profesores Lukas Nellen, Luciano Díaz y Eduardo Murrieta (publicación privada).

En el servidor MDS, se identificó como `/dev/sda` el disco con RAID 10 descrito en la sección anterior y se formateó como MDT (ver 1.3) con el siguiente comando:

```
# mkfs.lustre --mgs --mdt --backfstype=ldiskfs /
--mgsnode=172.16.1.200@tcp0 / --fsname=lustre /
--servicenode=172.16.1.200@tcp0 --comment="MGS+MDT" /
--verbose /dev/sda
```

donde 172.16.1.200 es la IP del servidor MDS.

Uno de los módulos de kernel de Lustre requiere los siguientes parámetros en el archivo `/etc/modprobe.d/lnet.conf`:

```
options lnet networks=tcp0(eno1)
```

donde `eno1` es la interfaz de red correspondiente a la conexión del MDS al clúster.

Se creó el archivo /etc/ldev.conf con la línea:

```
mds.lanot.unam.mx - MDT
/dev/disk/by-label/lustre-MDT0000
```

Se agregaron las siguientes líneas a las reglas del firewall con iptables:

```
-A INPUT -p tcp -m state --state NEW -m tcp --dport 22 -j ACCEPT
-A INPUT -m state --state NEW -m tcp -p tcp -s 172.16.0.0/16 -j ACCEPT
-A INPUT -m state --state NEW -m udp -p udp -s 172.16.0.0/16 -j ACCEPT
```

y desde la terminal se inició el sistema con el comando:

```
service  lustre start
```

En los servidores OSS, desde su sistema operativo, los discos de los gabinetes 4U se ven como 10 discos de 50TB cada uno.

```
# fdisk -l
...
Disk /dev/sda: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdb: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdc: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdd: 43980.5 GB, 43980462489600 bytes, 85899340800 sectors
Disk /dev/sde: 43084.4 GB, 43084431753216 bytes, 84149280768 sectors
Disk /dev/sdf: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdg: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdh: 54975.6 GB, 54975580864512 bytes, 107374181376 sectors
Disk /dev/sdi: 43980.5 GB, 43980462489600 bytes, 85899340800 sectors
Disk /dev/sdj: 43084.4 GB, 43084431753216 bytes, 84149280768 sectors
```

Se formateó cada uno como OST con el comando:

```
mkfs.lustre --ost --index=[0-9] --backfstype=ldiskfs \
--mgsnode=172.16.1.200@tcp0 --fsname=lustre \
--servicenode=172.16.1.1@tcp0 --comment="OST0[0-9]" \
--verbose /dev/sd[a-j]
```

donde 172.16.1.1 es la IP del OSS01. Para el OSS02 se usa el mismo comando, pero la IP se sustituye por 172.16.1.2 y los números van del 10 al 19 en los parámetros `index` y OST.

Para ambos OSS se creó el archivo `/etc/modprobe.d/lnet.conf` con la misma línea que en el MDS.

En el OSS01 se creó el archivo /etc/ldev.conf con las líneas:

```
oss01.lanot.unam.mx - OST00 /dev/disk/by-label/lustre-OST0000
oss01.lanot.unam.mx - OST01 /dev/disk/by-label/lustre-OST0001
oss01.lanot.unam.mx - OST02 /dev/disk/by-label/lustre-OST0002
oss01.lanot.unam.mx - OST03 /dev/disk/by-label/lustre-OST0003
```

```
oss01.lanot.unam.mx - OST04 /dev/disk/by-label/lustre-OST0004
oss01.lanot.unam.mx - OST05 /dev/disk/by-label/lustre-OST0005
oss01.lanot.unam.mx - OST06 /dev/disk/by-label/lustre-OST0006
oss01.lanot.unam.mx - OST07 /dev/disk/by-label/lustre-OST0007
oss01.lanot.unam.mx - OST08 /dev/disk/by-label/lustre-OST0008
oss01.lanot.unam.mx - OST09 /dev/disk/by-label/lustre-OST0009
```

Y en el OSS02 el archivo /etc/ldev.conf se creó con las líneas:

```
oss02.lanot.unam.mx - OST10 /dev/disk/by-label/lustre-OST000a
oss02.lanot.unam.mx - OST11 /dev/disk/by-label/lustre-OST000b
oss02.lanot.unam.mx - OST12 /dev/disk/by-label/lustre-OST000c
oss02.lanot.unam.mx - OST13 /dev/disk/by-label/lustre-OST000d
oss02.lanot.unam.mx - OST14 /dev/disk/by-label/lustre-OST000e
oss02.lanot.unam.mx - OST15 /dev/disk/by-label/lustre-OST000f
oss02.lanot.unam.mx - OST16 /dev/disk/by-label/lustre-OST0010
oss02.lanot.unam.mx - OST17 /dev/disk/by-label/lustre-OST0011
oss02.lanot.unam.mx - OST18 /dev/disk/by-label/lustre-OST0012
oss02.lanot.unam.mx - OST19 /dev/disk/by-label/lustre-OST0013
```

Y en ambos servidores se configuró el firewall como en el MDS y se inició el servicio con la línea:

```
service lustre start
```

Cada servidor OSS que se añada al sistema de almacenamiento, deberá configurarse de la misma manera.

2.3.2. Clientes de almacenamiento

En los servidores cliente del sistema de almacenamiento se deben instalar los paquetes de Lustre correspondientes (ver detalles en el apéndice B) pero la configuración es más sencilla. Como en la sección anterior, es necesario crear el archivo /etc/modprobe.d/lnet.conf que indique la interfaz de red conectada a la red del sistema de almacenamiento.

Para poder montar el disco del sistema de archivos Lustre, el cliente usa el mismo archivo de configuración del sistema operativo Linux para montar discos, `/etc/fstab` (File System Table). El sistema de archivos distribuido Lustre puede montarse como un disco virtual accesible con una configuiración similar a la de un disco remoto NFS:

```
172.16.1.200@tcp0:/lustre /depot lustre  defaults,_netdev 0 0
```

Donde `172.16.1.200` es la IP del MDS, `/depot` es el punto de montaje local y `lustre` el tipo de sistema de archivos. El parámetro `_netdev` previene que el sistema intente montarlo antes de que el servicio de red esté disponible.

Cada cliente que tenga acceso al sistema de almacenamiento, contará con un disco de trabajo local para hacer su procesamiento, que llamaremos `/data`. El

disco remoto, virtual, del sistema distribuido de almacenamiento, debe montarse en un punto de montaje local, que en nuestro caso llamaremos `/depot`.

El principal objetivo del sistema de almacenamiento es guardar la información que llega y se procesa diaramente al LANOT sin tener que eliminarla cada 12 horas. Tanto la tarea de guardar los datos originales como la de generar productos y guardarlos, se hará automáticamente mediante programas y scripts que se ejecuten periódicamente usando la herramienta `crontab` del sistema Linux. Se requiere también un usuario virtual que ejecute estas tareas, por lo cual se creó el usuario `lanotadm`, con derechos tanto de consulta como de escritura. Todos los programas y scripts ejecutables se guardarán en su carpeta `bin/` en su propio directorio de usuario.

Los datos que llegan continuamente serán leídos directamente de los servidores de adquisición del GOES16 y de los satélites polares y almacenados en la carpeta `/data/input`. Todos los trabajos de procesamiento se realizarán en carpetas temporales en el directorio `/var/tmp` y los resultados se guardarán en la carpeta `/data/output`.

A media noche se guardarán en `/depot` todos los conjuntos de datos que hayan sido adquiridos y procesados durante las últimas 24 horas, en carpetas organizadas por año y por semana, de acuerdo a las recomendaciones para sistemas Lustre, como no tener más de mil archivos en una sola carpeta y que dichos archivos no sean demasiado pequeños, lo recomendado es un tamaño cercano a 1GB.

2.4. Seguridad y Monitoreo

El cluster está aislado de la red externa y los servidores con acceso a la red externa cuentan con un firewall que restringe el acceso. Los datos se reciben continuamente, de día y de noche, por lo que conviene incluir mecanismos para monitoreo automático. A nivel hardware, lo más básico es prevenir fallas eléctricas, como apagones y saltos de voltaje, lo que se logra con un protector eléctrico con batería (*uninterruptible power supply* o UPS) que a su vez está conectado a una red con generador en el edificio, en caso de apagones.

Interrupción eléctrica Los servidores principal y secundario monitorean contínuamente el estado de las UPS por medio del demonio `apcupsd`. De esta manera se registra cuando la UPS entra en modo batería, debido a una falla o interrupción en la red eléctrica. Cuando la batería está a punto de agotarse, el servidor principal `cirrus` envía la señal a los demás servidores del cluster para que se apaguen de manera controlada y no abruptamente por interrupción eléctrica, lo que evita posibles daños al equipo y a la configuración.

Switch El switch tiene su propia IP y se puede obtener información sobre su estado desde los servidores, con el protocolo http.

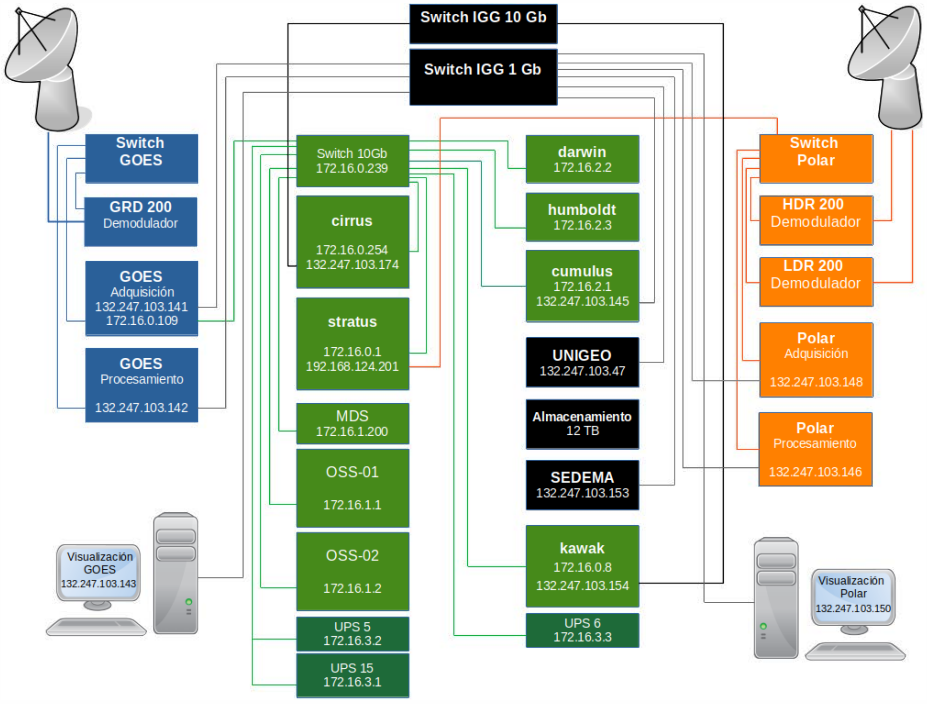

Figura 2.3: Sistema de almacenamiento integrado en la red local del LANOT. En verde el sistema de almacenamiento, los servidores principales `cirrus` y `stratus` y los servidores de servicio al usuario final, `kawak` y `cumulus`, unidos por el switch de 10Gb. En azul, la estación de recepción GOES y en anaranjado la estación de recepción polar. En negro los switch de la red del Instituto de Geografía y los servidores que comparten espacio en el rack pero no son parte del proyecto ni de la red local.

Discos duros Se prevee que cuando cumplan su vida media, los discos duros empezarán a fallar y conviene contar con un sistema que identifique rápidamente si algún disco está fallando y debe ser reemplazado. Cada cajón fue configurado en un arreglo RAID 6 con dos discos de paridad y uno de repuesto, lo que previene que el sistema se interrumpa aunque fallen físicamente dos discos. Para monitorear el estado físico de los discos dentro de uno de los arreglos, usamos la herramienta de la tarjeta controladora MegaRAID. Por ejemplo la siguiente instrucción desde el servidor maestro `cirrus` nos informará sobre el estado de los discos del OSS02:

```
$ sudo ssh oss02 /opt/MegaRAID/storcli/storcli64 /c0 show
```

Correo local Todas las operaciones programadas con `crontab` para realizar operaciones periódicas, del tipo de recuperar los datos de adquisición, procesarlos y guardarlos en el sistema Lustre, se documentan en correo electrónico local del usuario `lanotadm` y en cualquier momento se pueden consultar para verificar la correcta operación del sistema.

Capítulo 3

Resultados

3.1. Aplicaciones

En este capítulo se reporta el funcionamiento del sistema después de varios meses de su implementación. El satélite GOES 16[NN17] fue puesto en operación a finales de 2017, en la longitud 75W, lo que permite la observación constante de la mayor parte del hemisferio occidental. En el LANOT se reciben datos de sus sensores ABI (Advanced Baseline Imager), GLM (Geostationary Lightning Mapper) y los sensores solares EXIS y SUVI. El sensor ABI percibe 16 bandas en visible, infrarrojo cercano e infrarrojo térmico y genera datos de nivel **L1b** en disco completo (Full Disk o **FD**) cada 10 minutos, Estados Unidos continental (**CONUS**) cada 5 minutos y mesoescala (ubicación variable) cada minuto.

Con el paquete de código abierto CSPP (Community Satellite Processing Package) [Mar+14], se procesan las imágenes de nivel L1b y se generan 29 productos de nivel **L2** cada 10 minutos. Finalmente, el sensor de rayos GLM produce datos cada 20 segundos y se almacenan en conjuntos empaquetados cada hora. Desde junio de 2019 también se están almacenando los datos que se reciben de satélites de órbita polar, principalmente Suomi NPP y NOAA-20. Ese tipo de satélites recorren todo el planeta pero solamente dejan datos una o dos veces al día, cuando pasan en el rango de alcance de nuestra antena. Aunque su frecuencia de adquisición es mucho menor a la de un satélite de órbita geoestacionaria, la ventaja es que al abarcar un área menor, generan imágenes de mucho mayor resolución.

Los conjuntos de datos relacionados entre sí se guardan en archivos empaquetados con la herramienta `tar` y para ahorrar más espacio, en algunos casos se comprimen dichos paquetes. Las operaciones de almacenamiento se realizan automáticamente todos los días. Almacenando estos datos desde enero de 2019 a la fecha (marzo de 2020), se ha alcanzado poco más del 25 % del espacio disponible en el sistema de almacenamiento.

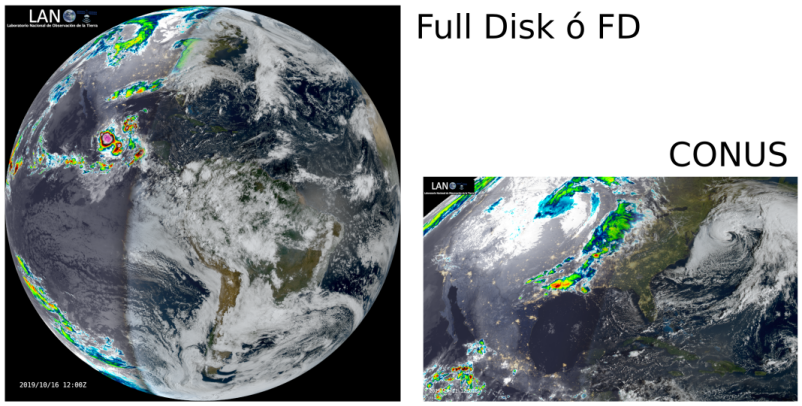

Figura 3.1: Regiones percibidas por el satélite GOES 16.

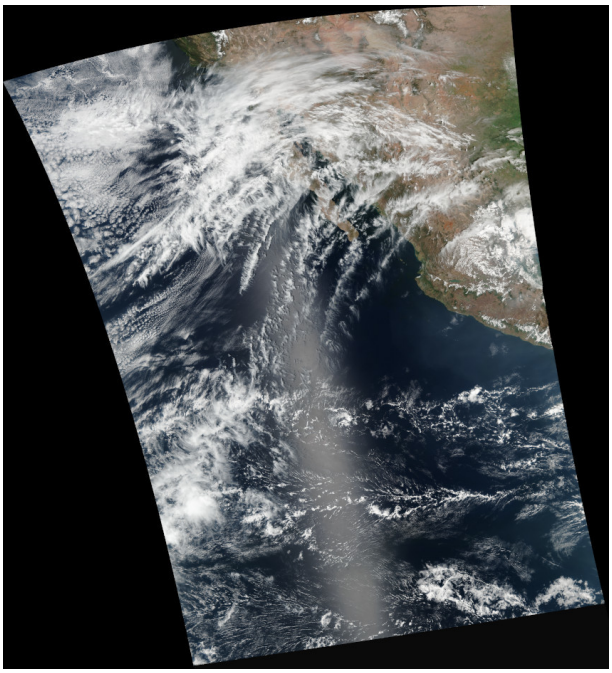

Figura 3.2: Región que captura el satélite de órbita polar Suomi NPP al pasar por la antena del LANOT.

Actualmente solo estamos respaldando los dos niveles en las regiones FD y CONUS que en conjunto generan diariamente alrededor de 20 mil archivos, por lo que su localización posterior podría ser muy difícil, a menos que se cuente con un buen sistema de organización. También almacenamos los tres niveles de los satélites polares, pero por su baja frecuencia, ocupan mucho menos espacio.

El árbol de directorios se organiza de acuerdo al satélite, sensor, nivel y fecha de adquisición. Para el sensor ABI del GOES-16, los paquetes de nivel L1b se guardan en la carpeta `/depot/goes16/abi/l1b/ [conus fd]` y los de nivel L2 en `/depot/goes16/abi/l2/ [conus fd] y` se distribuyen en carpetas de acuerdo al año y la semana del año, `yyyy/ww`, por lo que al final del año habrá 52 carpetas, una para cada semana. La semana del año se calcula dividiendo el día juliano (en este contexto, el número de día a partir del primero de enero) entre siete más uno.

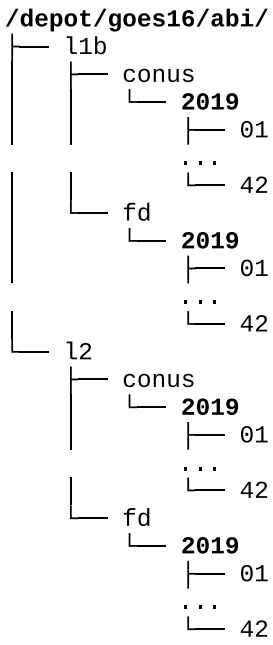

Figura 3.3: Árbol de directorios del sistema de almacenamiento del LANOT para el sensor ABI del satélite GOES-16.

Esta organización y nombrado de directorios ha facilitado la consulta de los datos tanto para el usuario final como para un proceso automático que se implemente a futuro, para dar acceso por medio de una interfaz web sin necesidad de que un usuario se conecte directamente al servidor.

Entre las aplicaciones que ya están haciendo uso del sistema de almacenamiento, se encuentra el de detección de incendios en México, coordinado por la

Dra. Lilia Manzo del Instituto de Geografía de la UNAM [Man+20; MGH19], en el que se usan las bandas 7 y 14 del GOES-16/ABI para detectar puntos de calor asociados con incendios forestales y se generan reportes cada 10 minutos. Un buen sistema de detección en tiempo casi real, es indispensable para para ubicar incendios y dispersión de nubes de humo para emitir alertas tempranas a los organismos encargados de controlar incendios antes de que sean graves, prevenir desastres, evaluar daños, calcular las emisiones de gases de invernadero y dar seguimiento al proceso de repoblación vegetal.

Otra aplicación del sistema de almacenamiento consiste en un repositorio para la Dirección General de Repositorios Universitarios (DGRU) [Men20], que guarda un catálogo de datos satelitales de última generación de los principales eventos meteorológicos ocurridos a partir de 2017 en México, incluyendo ciclones tropicales y extratropicales, puntos de calor, tormentas eléctricas y actividad volcánica para su consulta, descarga y visualización dinámica por medio de herramientas web. Estos productos son generados a partir de información proveniente principalmente del satélite GOES-16.

3.2. Cambios y adaptaciones imprevistas

Una mañana de junio de 2019, hubo un apagón imprevisto que duró cinco horas. La batería de la UPS se agotó y todo el sistema se vino abajo. Cuando se recuperó el suministro eléctrico, los servidores se pudieron echar a andar, salvo el OSS02. Cuando se hizo la revisión técnica correspondiente, se determinó que el backplane se había dañado y habría que reclamar la garantía. También se arruinó una de las tarjetas de red de 10 Gb de uno de los servidores de procesamiento. Como los apagones de la UNAM suelen durar pocos minutos y la batería del UPS principal tiene una duración mayor a las dos horas, no se había previsto esta eventualidad que provocó que los equipos se apagaran en desorden sin un apropiado *shutdown*.

A raíz de esa experiencia, se implementó un sistema de alerta, con la herramienta `apcupsd` que avisa a los servidores cuando la batería está por agotarse y los equipos se apagan en orden.

En enero de 2020, el robo de un tablero eléctrico en el edificio principal del Instituto, provocó una falla eléctrica de varias horas. Gracias al sistema de alerta, no hubo ninguna pérdida material que lamentar y solamente se perdieron los datos que se habrían recibido en las estaciones durante el tiempo que duró el apagón. Al volver la electricidad, los servidores volvieron a sus funciones sin ningún problema.

Salvo ese caso, en año y medio no ha habido otras situaciones imprevistas que nos hayan obligado a revisar el diseño original. A finales de 2019 empezaron a fallar algunos discos de los OSTs y fueron oportunamente reemplazados por los discos que previsoramente se adquirieron para ese fin.

Capítulo 4

Conclusión y posibles mejoras futuras

Mediante un sistema de código abierto, hardware armado bajo diseño y configuración propia, fue posible implementar un sistema de almacenamiento para el LANOT a un costo de cerca del 8 % de lo que se había cotizado con soluciones comerciales cerradas.

Aparte del considerable ahorro, se cuenta con el conocimiento práctico para administrar este sistema sin necesidad de servicios de soporte externos, de los que se dependería de haber adquirido una solución comercial cerrada como caja negra. Los servicios de soporte comerciales, en muchos casos son expertos en las necesidades de compañías privadas con fines de lucro pero desconocen las necesidades de un laboratorio con fines académicos y de servicio público. En el mismo LANOT actualmente se sufre esa situación, con otro sistema cerrado adquirido previamente.

Además del ahorro y el acceso al conocimiento, la solución elegida, a más de un año de su implementacion, superó pruebas difíciles y ha cumplido las necesidades por las que fue creada, además de ampliar los servicios que puede ofrecer el LANOT. Es una solución exitosa y un ejemplo de que vale la pena impulsar el desarrollo y la innovación dentro de la universidad y de la nación, en lugar de depender de soluciones importadas, comerciales y cerradas, aunque sean aparentemente «probadas».

Para llegar a esta solución, se aprovechó la experiencia de 7 años del proyecto HAWC [A S15], en el Instituto de Ciencias Nucleares de la UNAM; lo cual demuestra que un proyecto de ciencia básica puede generar innovación y beneficios colaterales a sus objetivos.

Siendo un sistema vivo en operación, seguirá siendo adaptado y mejorado a las necesidades de los usuarios del LANOT. Ofrecemos nuestra experiencia y conocimientos para colaborar en el futuro con proyectos similares.

Bibliografía

[A S15] M. M. González A. Sandoval. *Laboratorio Nacional HAWC de Rayos Gamma*. 2015. URL: http://labunam.unam.mx/micrositio/HAWC/.

[A U12] et al A. U. Abeysekara. "On the sensitivity of the HAWC observatory to gamma-ray bursts". En: *Astroparticle Physics* 35 (10 2012), págs. 641-650. DOI: 10.1016/j.astropartphys.2012.02.001.

[Agu18] Raúl Aguirre. "Laboratorio Nacional de Observación de la Tierra (LANOT)". En: *Investigaciones Geográficas* 96 (2018). DOI: dx.doi.org/10.14350/rig.59730.

[Bra18] Peter J. Braam. *The Lustre filesystem*. 2018. URL: http://lustre.org/.

[Ear] Group on Earth Observations. *About GEONETCast*. Ed. por Group on Earth Observations. URL: https://www.earthobservations.org/geonetcast.php.

[Man+20] Lilia Manzo y col. "Detection of vegetation fires in Mexico using GOES-16/ABI images: Algorithm description and preliminary assessment". 2020.

[Mar+14] Graeme Martin y col. "Introducing CSPP GEO: A Geostationary Satellite Data Processing Package for Direct Broadcast Users". En: 2014.

[Men20] Uriel Mendoza. "Repositorio interactivo de datos satelitales de eventos meteorológicos del Laboratorio Nacional de Observación de la Tierra". 2020.

[MGH19] Lilia Manzo, C. Gómez y J. Hernández. "Monitoreo avanzado de incendios forestales y quemas agropecuarias: una estrategia para reducir el riego de desastre". En: *Memorias del 1er. Encuentro Multisectorial hacia la Gestión Integral del riesgo de desastre (GIRD): Construyendo la Política Nacional* (Ciudad de México). 24 de oct. de 2019.

[MP17] Enrique Murrieta y Enrique Palacios. "Tecnologias para el almacenamiento de grandes volúmenes de datos en HPC LUSTRE". En: *Taller ISUM 2017*. Instituto de Ciencias Nucleares, UNAM. 2017.

[NN17] NOAA y NASA. *Página oficial del programa GOES-R*. 2017. URL: https://www.goes-r.gov/.

[P H19] A. Carramiñana P. Hüntemeyer A. Sandoval. *HAWC: the High-Altitude Water Cherenkov Observatory*. 2011-2019. URL: https://www.hawc-observatory.org/.

Apéndice A

Breve Manual de Operación

Instrucciones breves para la operación básica del sistema y qué hacer en casos en que sea necesario apagar todo el cluster debido a mantenimiento en la red eléctrica, mudanza del centro de datos o cualquier caso similar. Las operaciones de administración del cluster se ejecutan desde el servidor maestro `cirrus` y en caso de falla de este último, en el servidor secundario `stratus`. Por seguridad se recomienda usar una cuenta de usuario con derechos administrativos (`sudo`) y nunca directamente desde el usuario `root`.

A.1. Iniciar el sistema Lustre

```
sudo ssh mds service lustre start
sudo ssh oss01 service lustre start
sudo ssh oss02 service lustre start
sudo mount /depot
```

A.2. Desactivar el sistema Lustre

Primero detener todos los procesos que accesan /depot, que en nuestro caso están programados con el servicio `cron` y desmontar /depot en todos los servidores.

```
sudo service cron stop
sudo ssh stratus service cron stop
sudo umount /depot
sudo ssh stratus umount /depot
```

Posteriormente detener el sistema Lustre en todos los servidores, empezando por el MDS.

```
sudo ssh mds service lustre stop
```

```
sudo ssh oss01 service lustre stop
sudo ssh oss02 service lustre stop
```

Para apagar en orden los servidores del sistema, apagar primero el sistema de almacenamiento y luego los servidores de procesamiento:

```
sudo ssh mds shutdown -h now
sudo ssh oss01 shutdown -h now
sudo ssh oss02 shutdown -h now
sudo ssh stratus shutdown -h now
sudo shutdown -h now
```

A.3. Monitorear

Para consultar el estado de los discos físicos, usar el comando storcli64 para la tarjeta controladora Megaraid instalada en el sistema.

```
sudo ssh oss01 /opt/MegaRAID/storcli/storcli64 /c0 show
sudo ssh oss02 /opt/MegaRAID/storcli/storcli64 /c0 show
sudo ssh oss02 /opt/MegaRAID/storcli/storcli64 /c1 show
```

Poner atención en la tabla TOPOLOGY y que en la columna State todos los discos tengan el estado `Onln` (en línea).

Para consultar la temperatura de ambiente en el sitio, el estado de la batería y otras variables del UPS, usar el comando `apcaccess` y poner atención en las líneas:

```
STATUS   : ONLINE
BCHARGE  : 100.0 Percent
TIMELEFT : 123.0 Minutes
ITEMP    : 21.0 C
ALARMDEL : No alarm
```

Una carga de batería menor al 10 % y un tiempo menor de 10 minutos, así como una temperatura mayor de 27 C, son señal de un problema físico en el centro de datos, probablemente relacionado con una falla eléctrica larga (de varias horas) o del sistema de enfriamiento.

Para verificar el estado de las tareas automáticas de respaldo y procesamiento de datos, consultar el correo electrónico local del usuario `lanotadm` con el comando `mutt` en ambos servidores cirrus y stratus.

Apéndice B

Paquetería del sistema operativo

En los servidores de almacenamiento descritos en el capítulo 2, se instaló el sistema operativo Linux con la distribución CentOS 7 y los paquetes básicos para un sistema en red y los paquetes necesarios para echar a andar un sistema Lustre, así como herramientas para el manejo de sistemas de archivos, módulos del kernel, etc. En nuestra implementación en particular se instalaron los siguientes paquetes:

```
e2fsprogs-1.44.3.wc1-0.el7.x86_64.rpm
e2fsprogs-libs-1.44.3.wc1-0.el7.x86_64.rpm
e2fsprogs-static-1.44.3.wc1-0.el7.x86_64.rpm
kernel-3.10.0-862.9.1.el7_lustre.x86_64.rpm
kernel-devel-3.10.0-862.9.1.el7_lustre.x86_64.rpm
kernel-firmware-2.6.32-573.12.1.el6_lustre.x86_64.rpm
kernel-headers-3.10.0-862.9.1.el7_lustre.x86_64.rpm
kmod-lustre-2.10.5-1.el7.x86_64.rpm
kmod-lustre-osd-ldiskfs-2.10.5-1.el7.x86_64.rpm
kmod-lustre-tests-2.10.5-1.el7.x86_64.rpm
libcom_err-1.44.3.wc1-0.el7.x86_64.rpm
libss-1.44.3.wc1-0.el7.x86_64.rpm
libuutil1-0.7.9-1.el7.x86_64.rpm
lustre-2.10.5-1.el7.x86_64.rpm
lustre-2.10.5-1.src.rpm
lustre-all-dkms-2.10.5-1.el7.noarch.rpm
lustre-all-dkms-2.10.5-1.el7.src.rpm
lustre-iokit-2.10.5-1.el7.x86_64.rpm
lustre-ldiskfs-dkms-2.10.5-1.el7.noarch.rpm
lustre-ldiskfs-dkms-2.10.5-1.el7.src.rpm
lustre-modules-2.8.0-2.6.32_573.12.1.el6_lustre.x86_64.x86_64.rpm
lustre-osd-ldiskfs-mount-2.10.5-1.el7.x86_64.rpm
```

lustre-source-2.8.0-2.6.32_573.12.1.el6_lustre.x86_64.x86_64.rpm
lustre-tests-2.10.5-1.el7.x86_64.rpm
perf-3.10.0-862.9.1.el7_lustre.x86_64.rpm
python-perf-2.6.32-573.12.1.el6_lustre.x86_64.rpm

En el servidor maestro de correspondientes de administración y procesamiento, `cirrus`, que a su vez será el cliente principal del sistema de almacenamiento, se instalaron los siguientes paquetes:

kmod-lustre-client-2.10.5-1.el7.x86_64.rpm
kmod-lustre-client-tests-2.10.5-1.el7.x86_64.rpm
lustre-2.10.5-1.src.rpm
lustre-client-2.10.5-1.el7.x86_64.rpm
lustre-client-debuginfo-2.10.5-1.el7.x86_64.rpm
lustre-client-tests-2.10.5-1.el7.x86_64.rpm
lustre-iokit-2.10.5-1.el7.x86_64.rpm

En el servidor secundario `stratus` se instaló el sistema operativo Linux Debian 9.9 con el kernel 4.9 pues aunque la versión actual estable de Debian tiene un kernel más reciente, no es compatible con la versión de Lustre que instalamos. Se compilaron los módulos del kernel y se instalaron herramientas auxiliares, con los paquetes:

lustre-client-modules-4.9.168_2.10.8-1_amd64.deb
lustre-utils.