

LUSTRE PARALLEL FILE SYSTEM EN AWS

**TREBALL FI DE GRAU
2n ASIXc**

**Adrian Manzanares
Oriol Coll
Pau Ruiz
Victor Gonzalez**

**Institut Tecnològic de Barcelona
Tutor: Victor Marquina de Castro
Curs 2019-2020**

INDEX

1. Introducció	3
2. Parallel File System	4
2.1 Operacions d'entrada/sortida per segon (IOPS)	7
3. Softwares Parallel File System	8
3.1 General Parallel File System (GPFS IBM)	8
3.2 BeeGFS	10
3.3 Lustre	12
4. Elecció del software	15
4.1 Components Lustre	15
4.2 Xarxes Lustre	16
4.3 Esquema del funcionament de Lustre	16
5. Amazon Web Services (AWS)	18
5.1 Configuració de la xarxa en AWS	18
5.2 Creació de la instància en AWS	20
6. Instal·lació de Lustre en CentOS 7 amb AWS	22
6.1 Creació part comuna nodes	23
6.2 Instal·lació dels repositoris dels servidors	24
6.3 Mòdul LNET als servidors (MGS-MDS i OSS)	26
6.4 Instal·lació dels repositoris del Client	27
6.5 Activació del mòdul Lustre al Client	28
6.6 Muntatge del Lustre File System	29
6.7 Comprovació del Sistema de fitxers Lustre	30
7. Script d'instal·lació de Lustre	31
8. Conclusió	32
9. Glossari	33
10. Webgrafia	34

1. Introducció

Aquest document és el projecte realitzat per Oriol Coll, Pau Ruiz, Adrián Manzanares i Víctor González, estudiants del grau superior d'Administració de Sistemes en la Xarxa perfil Ciberseguretat. S'ha realitzat amb el suport del tutor Víctor Marquina i amb l'ajuda de l'empresa HPCNow!. Aquest treball s'ha fet durant el període de maig i juny de 2020.

Hem escollit aquest projecte com una alternativa a l'anterior proposta, el qual era muntar un Parallel File System en més d'una Raspberry Pi. Tot i això aquesta nova proposta té bastants similituds amb l'anterior, que degut al confinament no s'ha pogut desenvolupar.

Aquest projecte es basa en una disciplina gairebé desconeguda per tots nosaltres, la supercomputació, i aquest ha sigut un dels principals motius pels quals ens ha interessat desenvolupar-lo amb totes les nostres capacitats.

Per poder crear el Parallel File System (PFS) primer hem investigat per poder entendre els seus principals conceptes, com funciona i els dispositius necessaris per poder funcionar, una vegada fet això hem fet un petit estudi de mercat de tres tipus de software disponibles de PFS i hem escollit un entre els tres, en aquest cas Lustre, després hem configurat la infraestructura d'aquest PFS de Lustre amb Amazon Web Services, i finalment hem instal·lat el programari Lustre en aquesta infraestructura i quan ha funcionat tot correctament, hem scriptat tot el procés.

2. Parallel File System

Abans d'aprofundir en un **sistema de fitxers paral·lels** haurem d'establir els conceptes generals i definicions que formen aquest sistema, ja que aquest és una especialització de dos conceptes en concret.

El primer concepte és un **sistema de fitxers**, aquest anomenat també **File System (FS)**, el qual permet controlar com s'emmagatzema i es recuperen les dades d'un sistema.

Aquest sistema de fitxers és molt important, ja que sense ell les dades dels sistemes no tindrien cap ordre i estarien barrejades entre si, una gran massa de dades sense forma, per aquesta raó el sistema de fitxers s'encarrega de separar les dades en parts i indicar-li el seu nom corresponent, d'aquesta manera aconseguim que aquestes dades tinguin una forma i s'identifiquin, el qual dona lloc el nom de fitxer, en estructurar, administrar i afegir regles lògiques per administrar els grups de dades i noms aconseguim aquest sistema de fitxers.

Cada tipus de sistema de fitxers té la seva pròpia lògica i estructura, propietats, flexibilitat, seguretat, mida i altres consideracions...

Els sistemes de fitxers es poden utilitzar tant en dispositius d'emmagatzematge (SSD, HDD), a través de protocols de xarxa (NFS, SMB) o virtuals (procfs, sysfs).

El sistema d'arxius administra l'accés tant al contingut dels arxius com les metadades sobre aquests arxius, és responsable d'organitzar l'espai d'emmagatzematge; la fiabilitat, l'eficiència i l'ajust pel que fa a l'emmagatzematge físic són aspectes rellevants pel disseny.

El segon concepte és un **sistema de fitxers en clúster**, que és un sistema de fitxers el qual es comparteix en muntar-se simultàniament en diversos servidors, nodes i ofereix unes característiques com, adreçament independent de la ubicació i redundància, el qual millora la confiança, disponibilitat i redueix la complexitat de les altres parts del clúster.

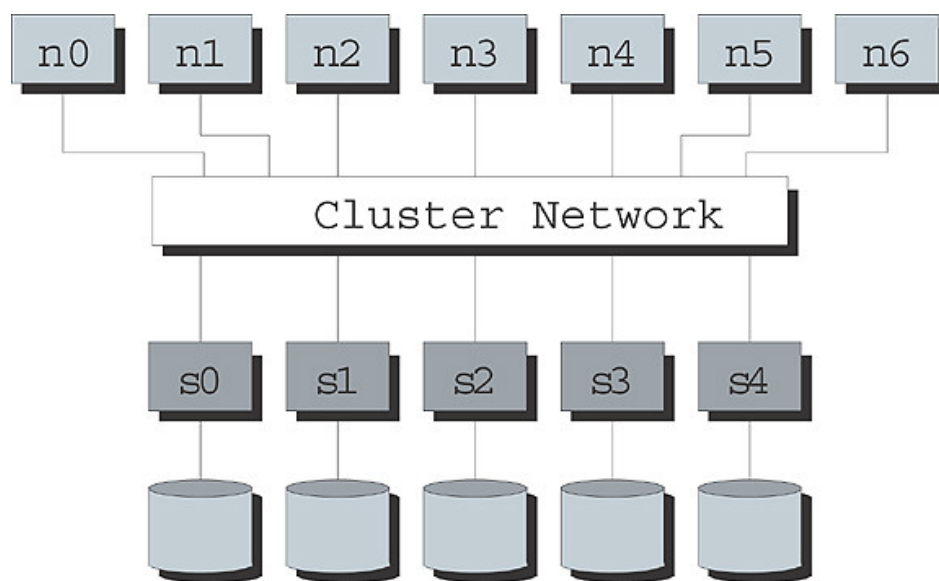
Hi ha dos tipus de sistema de fitxers en clúster, el **sistema de fitxers de disc compartit** i el **sistema de fitxers distribuïts**.

El **sistema de fitxers de disc compartit** utilitza una xarxa d'emmagatzematge (SAN) el qual permet que diversos dispositius obtinguin accés directe al disc o el registre físic. Aquest tipus de sistema afegeix mecanismes com el control de la concurrència, una vista coherent i serialitzat, d'aquesta manera es pot prevenir la corrupció i la pèrdua involuntària de dades, utilitzant el "**Fencing**", el qual és el procés per aïllar un node d'un clúster quan aquest funciona incorrectament.

El **sistema de fitxers distribuïts**, no comparteixen a l'accés directe al disc o al registre físic com el de disc compartit, però utilitza un protocol de xarxa. Aquests poden restringir l'accés del sistema mitjançant "acces list" tant en servidors com en clients.

El sistema de fitxer en paral·lel es basa amb aquest tipus de sistema però hi ha una petita diferència entre aquests dos, mentre que el paral·lel les dades poden estar a diversos dispositius alhora sense dependre de cap dispositiu de la xarxa, el sistema de fitxers distribuïts pot estar a diferents dispositius alhora però per recollir aquesta informació passa per un únic dispositiu i es forma un coll d'ampolla, el qual amb el paral·lel no.

El **sistema de fitxer paral·lel** és un tipus de sistema de fitxers en clúster, el qual distribueix les dades a través de múltiples nodes d'emmagatzematge, generalment per afegir redundància i augmentar el seu rendiment, ja que principalment estan dissenyats per oferir un servei d'alt rendiment com HPC.



Esquema del sistema d'un clúster

La seva funció principal és que el conjunt de les dades es trenca i aquests blocs es distribueixen/seccionen en múltiples dispositius d'emmagatzematge. Utilitza un espai de noms globals per facilitar l'accés de les dades, normalment utilitza servidors dedicats a l'emmagatzematge de les metadades i la lectura/escriptura d'aquestes dades en múltiples dispositius d'emmagatzematge i d'aquesta manera utilitzar múltiples rutes de E/S simultànies, per més redundància i rendiment.

La E/S paral·lela és el compliment de múltiples operacions d'entrada/sortida al mateix moment, és a dir, sortides simultànies a dispositius d'emmagatzematge o virtualització.

Característiques rellevants en un Paral·lel File System:

- **Banda ampla de E/S:** Per aquest tipus de sistema les sol·licituds de E/S realitzades als servidors del sistema d'arxius tenen quasi la mateixa mida que les sol·licituds de E/S de l'aplicació.
- **Metadades:** Els sistemes d'arxius paral·lels es van dissenyar per admetre milers de milions d'arxius amb accés a metadades d'alt rendiment com a part del disseny original. Han de tenir les seves metadades amb sincronització i també compatibles amb POSIX (Interfície d'un sistema operatiu portàtil estandarditzat). Un exemple és que admeten fer 30,000 trucades stat() per segon al client.
- **Escalabilitat:** La gran majoria de PFS tenen aquesta característica de suport perquè estan dissenyats d'aquesta manera, i així poder tenir un creixement en la xarxa (més dispositius, servidors, etc.) sense que afecti la qualitat de tot el sistema.
- **Estructura de la xarxa:** Aquesta es construeix en agrupar diversos components diferents, generalment múltiples servidors, múltiples discos, operant junts per proporcionar un espai de noms unificat, el qual ofereix millor rendiment. El client no coneix els límits físics que conformen el sistema d'arxius.

2.1 Operacions d'entrada/sortida per segon (IOPS)

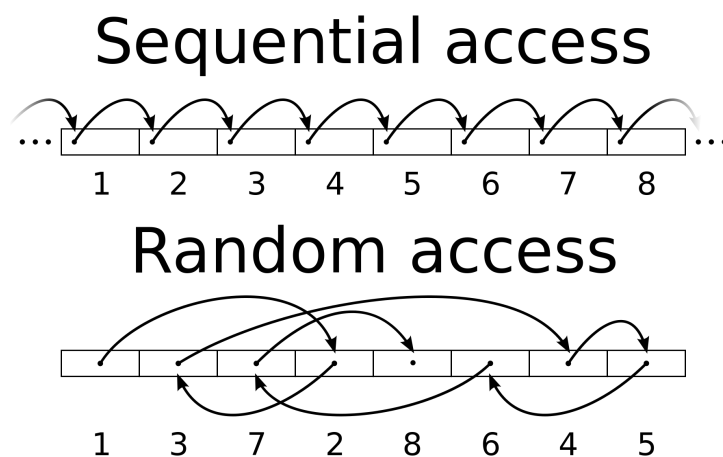
És una mesura de rendiment d'entrada/sortida utilitzada per definir dispositius d'emmagatzematge com, HDD, SSD i SAN.

Per definir les característiques de rendiment d'aquests dispositius d'emmagatzematge, es necessita un mínim de 3 mètriques simultànies, IOPS, temps de resposta i càrrega de treball (en aplicació).

Aquest número específic variar depenent de diverses variables com el balanç d'operacions de lectura/escriptura, la combinació de patrons d'accés seqüencial i aleatori, el nombre de fils de treball, la profunditat de la cua i les mides d'aquests blocs de dades.

Operacions seqüencials, accedeixen a ubicacions en el dispositiu de manera contigua, normalment per a fitxers més grans.

Operacions aleatòries, accedeixen a ubicacions de manera no contigua, normalment per fitxers més petits.



Esquema del funcionament d'operacions seqüencials i aleatòries

3. Softwares Parallel File System

En aquest apartat explicarem les característiques dels diferents softwares que ofereix el mercat i explicarem quin hem escollit per realitzar la part pràctica del nostre projecte i el perquè de la nostra elecció. I aquests sistemes són:

- General Parallel File System
- BeeGFS
- Lustre

3.1 General Parallel File System (GPFS IBM)

General Parallel File System (GPFS IBM Spectrum Scale) és un sistema de fitxers distribuït d'alt rendiment desenvolupat per IBM, aquest està compost d'un conjunt d'ordinadors que formen un clúster GPFS. Alguns dels membres del clúster proporcionen els discos físics accessibles per tots els nodes de sistema. La inclusió i exclusió de membres del clúster pot realitzar-se en funcionament.

L'origen del GPFS és el sistema de fitxers Tiger Shark, un projecte del centre d'investigació Almaden d'IBM a principis de 1993, encara que la primera versió comercial de GPFS va ser publicada el 1998 per al sistema operatiu AIX.

Inicialment, GPFS es dissenya per suportar altes taxes de transferència que necessiten les aplicacions multimèdia i quan un node realitza una operació sobre el sistema de fitxers les dades es distribueixen en tires ("**striping**") i són emmagatzemades en diverses de les màquines que serveixen de discs.

GPFS proporciona accés simultani a arxius d'alta velocitat a aplicacions que s'executen en múltiples nodes de clústers. Es pot utilitzar amb clústers AIX, clústers Linux, en Microsoft Windows Server o un clúster heterogeni de nodes AIX, Linux i Windows que s'executen en arquitectures de processador x86, POWER o IBMz. A més de proporcionar capacitats d'emmagatzematge de sistema d'arxius, GPFS proporciona eines per a la gestió i administració del clúster GPFS i permet l'accés compartit als sistemes d'arxius dels clústers remots de GPFS.

Beneficis d'utilitzar el General Parallel File System (IBM Spectrum Scale):

- Arquitectura moderna.
- Un espai de noms global que pot estendre els centres de dades i el món.
- Accedir directament a les dades del fitxer o utilitzant NFS, SMB, Object i Hadoop.
- Funcionament i escala de capacitat gairebé sense límits.
- Col·locació automatitzada de dades i migració de dades.
- Aprofiteu l'emmagatzematge de mercaderies menys costos.
- Integrat amb IBM Spectrum Archive i IBM Spectrum Protect.
- Una llarga història de fiabilitat contrastada en diverses indústries.

Característiques clau del General Parallel File System (IBM Spectrum Scale):

- Proporciona una escalabilitat extrema.
- Redueix els costos d'emmagatzematge.
- Millora la seguretat i l'eficiència de gestió.
- Intercanvi de fitxers global amb gestió activa de fitxers.
- Transforma l'economia de les dades.
- Gestionar i obtenir informació sobre quantitats massives de dades.
- Disponible en Amazon Web Services.

Requeriments de maquinari:

Maquinari compatible: arquitectura x86: processadors Intel EM64T o AMD Opteron, memòria del sistema d'1 GB d'arquitectura IBM POWER: AIX v6.1 o v7.1, Linux a POWER3 (mínim), mínima memòria del sistema d'1 GB, i Z Systems (només a linux)

Especificacions tècniques:

Sistemes operatius compatibles: IBM AIX®; Linux: Red Hat, SUSE Linux Enterprise Server; Microsoft Windows Server 2012, Microsoft Windows 7; IBM z Systems

Protocols: POSIX, GPFS, NFS v4.0, SMB v3.0

Emmagatzematge d'objectes al núvol:

- Proveïdors compatibles amb IBM Cloud Storage System (Cleversafe), Amazon S3, IBM Cloud Native Object, OpenStack Swift i Amazon S3.
- Nombre màxim de fitxers / sistema de fitxers: 264 (9 quintilions) de fitxers per sistema de fitxers
- Mida màxima del sistema de fitxers: 8 Exabytes (EB)
- Nombre mínim / màxim de nodes: 1 - 16.384

3.2 BeeGFS

BeeGFS és un sistema de fitxers en paral·lel, desenvolupat i optimitzat per a informàtica en un nivell d'alt rendiment. BeeGFS inclou una arquitectura de metadades distribuïda per raons d'escalabilitat i flexibilitat. El seu aspecte més important és el rendiment de dades. BeeGFS és un programari lliure, és a dir, que es pot descarregar i utilitzar gratuïtament des del lloc web del projecte.

Aquest projecte va començar en 2005 i la primera versió beta va sortir l'any 2007, i des d'aquest punt fins ara se segueix actualitzant constantment amb noves versions. BeeGFS es va desenvolupar seguint els tres següents conceptes clau: escalabilitat, flexibilitat i bona usabilitat. Per fer funcionar-lo només es necessita qualsevol màquina Linux i consta de diversos components que inclouen serveis per a clients, servidors de metadades i servidors d'emmagatzematge. A més, té un servei per a l'amfitrió de gestió i un altre per a un sistema d'administració i control gràfic per fer un ús més còmode.

Per poder executar BeeGFS és necessària almenys una instància del servidor de metadades o del servidor d'emmagatzematge. Però BeeGFS permet múltiples instàncies per cada servei per distribuir la gran càrrega d'un gran nombre de clients alhora, i l'escalabilitat de cada component assegura que el mateix sistema sigui escalable.

BeeGFS gestiona la distribució del contingut dels fitxers distribuint-los per diversos servidors d'emmagatzematge utilitzant striping, és a dir, cada arxiu és distribuït en parts d'una determinada mida i aquestes parts són distribuïdes pels servidors d'emmagatzematge existents. La mida d'aquestes parts es determina per l'administrador del sistema de fitxers.

A més a més, també les metadades són distribuïdes per diversos servidors de metadades en un nivell de directoris, en el que cada servidor està emmagatzemant una part de l'arbre complet del sistema de fitxers. Això fa que sigui més ràpid l'accés a les dades del nostre servidor.

Altres punts a favor serien que els clients tant com els servidors de metadada o d'emmagatzematge poden ser afegits en un sistema sense tenir un temps d'inactivitat, el client ell mateix és un mòdul de kernel lleuger que no requereix cap patch del kernel. Els servidors funcionen a sobre d'un sistema de fitxers local existent, no hi ha restriccions al tipus de sistema de fitxers subjacents sempre que sigui compatible amb POSIX (Portable Operating System Interface uniX), es recomana utilitzar ext4 pels servidors de metadades i XFS pels servidors d'emmagatzematge.

Per últim una característica molt important de BeeGFS és que no hi ha un requisit estricte de hardware dedicat a serveis individuals, el seu disseny permet que l'administrador del sistema de fitxers comenci els serveis en qualsevol combinació de hardware amb possibilitat de millorar-lo en un futur.

En resum podem dir que BeeGFS és un programari de software lliure, fàcil d'utilitzar, ben optimitzat que estén les dades dels usuaris a diversos servidors de forma transparent. Si s'augmenta el nombre de servidors i discs del sistema, es pot simplement escalar el rendiment i la capacitat del sistema de fitxers al nivell que necessitem, des de petits clústers fins a sistemes de classe empresarial amb milers de nodes.

Avantatges d'utilitzar BeeGFS

- Desenvolupat amb l'objectiu de centrar-se en el màxim rendiment, escalabilitat, flexibilitat i robustesa.
- Basat en la interfície de sistema de fitxers POSIX, el que significa que les aplicacions no han de ser reescrites ni modificades.
- Separa les metadades dels chunks de fitxers d'usuari dels servidors.
- Es dirigeix a tothom, usuaris que necessiten un emmagatzematge gran o usuaris que necessiten un emmagatzematge ràpid.
- Programari de codi obert totalment gratuït.

3.3 Lustre

Lustre és un sistema d'arxius distribuït de codi obert que se sol utilitzar en clústers a gran escala. El sistema d'arxius Lustre és compatible amb Linux i dóna una interfície de sistema d'arxius UNIX basada en l'estàndard POSIX. Quan un programa està escrit en els estàndards POSIX podem estar segurs del fet que el podrem importar fàcilment entre una gran quantitats de derivats UNIX (entre ells Linux).

Lustre és notablement conegut per donar servei a clústers de supercomputació (HPC) amb moltíssims clients, petabytes d'informació i molts gigabytes per segon de rendiment.

Que ofereix y que no ofereix Lustre?

Respecte al que ofereix podem dir que els sistemes Lustre poden escalar molt notablement la capacitat i el rendiment per a qualsevol necessitat. Això fa que no sigui necessari utilitzar molts sistemes d'arxius separats o utilitzar un per cada grup d'ordinadors. Això simplifica les coses perquè evita que s'hagin de copiar dades entre clústers.

Pel que fa al que no ofereix o per a què camps no seria del tot òptim Lustre podríem dir que tot i que el sistema d'arxius Lustre és molt potent i pot funcionar en diversos entorns potser no és la millor opció en tots els àmbits. Lustre és adequat principalment per usos que sobrepassi la capacitat que pot donar un sol servidor.

En l'actualitat Lustre no és adequat per models "punt a punt" on client i servidor s'executen al mateix node a causa de la falta de replicació de dades en el software Lustre.

Característiques Lustre

Una de les característiques més a tenir en compte del software Lustre és la seva gran capacitat d'escalabilitat i rendiment. L'escalabilitat és la capacitat que té un sistema d'augmentar la seva càrrega de treball sense que es vegin afectats ni el rendiment ni la qualitat del servei. Tot i això s'ha de tenir en compte que l'escalabilitat i el rendiment depenen de l'amplada de banda disponible del disc i de la xarxa i de la potència de processament que tinguin els servidors.

- **Sistema d'arxius ext4:** Lustre utilitza per emmagatzemar dades i metadades un sistema d'arxius millorat del sistema d'arxius ext4. La versió té el nom de ldiskfs i s'ha millorat per proporcionar un més alt rendiment necessari per al sistema d'arxius Lustre.

A Lustre 2.4 es va introduir també la possibilitat d'utilitzar el sistema d'arxius ZFS. Això permet que Lustre utilitzi les funcions d'escalabilitat i integritat de dades que ZFS ofereix.

- **Estandardització POSIX:** Lustre presenta a tots els seus clients una semàntica POSIX estàndard i accés de lectura i escriptura a tots els objectes compartits. POSIX és una família d'estàndards utilitzada per aclarir i uniformar les interfícies de programació de les aplicacions (i els problemes auxiliars, com les utilitats Shell de la línia d'ordres) proporcionades per Unix. Quan un programa està escrit en els estàndards POSIX podem estar segurs de què el podrem importar fàcilment entre una gran quantitat de derivats UNIX (entre ells Linux).
- **Xarxes heterogènies d'alt rendiment:** Lustre admet diverses xarxes d'alt rendiment i baixa latència. A més també permet l'accés a memòria directa remota (RDMA) per InfiniBand. També es poden combinar diverses xarxes RDMA per mitjà de Lustre per obtenir el màxim rendiment. Lustre també inclou diagnòstic de xarxa.
- **Alta disponibilitat:** El sistema de fitxers Lustre permet la commutació basada en actiu/actiu mitjançant particions d'emmagatzematge compartit per orientacions OSS. El sistema Lustre pot funcionar amb diversos gestors d'alta disponibilitat per permetre la recuperació automàtica i no tenir un únic punt de fallada. Lustre ofereix també protecció de muntatge múltiple (MMP) proporciona protecció integrada dels errors dels sistemes altament disponibles que, d'altra manera, podrien causar la corrupció del sistema d'arxius.
- **Seguretat:** Per defecte, només es permeten les connexions TCP des de ports privilegiats.
- **Llista de control d'accés (ACL):** La seguretat de Lustre segueix un sistema de fitxers UNIX millorat amb ACL POSIX.
- **Interoperabilitat:** La interoperabilitat és la capacitat dels sistemes informàtics de compartir dades e informació entre els diferents procediments que suporten. El sistema de fitxers Luster funciona amb una gran varietat d'arquitectures de CPU i clústers i és interoperable.
- **Arquitectura basada en objecte:** Els clients estan aïllats de l'estructura de fitxers. Això permet actualitzar l'arquitectura d'emmagatzematge sense que el client es vegi afectat.
- **Fitxer de gran byte i bloqueig de metadades de gra fi:** Molts clients poden llegir i modificar el mateix fitxer o directori simultàniament. El gestor de bloqueig distribuït de Lustre (LDLM) garanteix que els fitxers siguin coherents entre tots els clients i servidors del sistema de fitxers. El MDL LDLM gestiona els bloquejos en els permisos i els noms de ruta dels nodes.
- **Quotes:** En el sistema de fitxers Lustre les quotes d'usuaris i grups estan disponibles.
- **Creixement de la capacitat:** La mida d'un sistema de fitxers Lustre i l'amplada de banda del clúster es poden augmentar sense interrupcions si afegim nous OST i MDT al clúster.

- **Protecció de la integritat de dades en xarxa:** Es fa un resum de totes les dades enviades dels clients a les OSS. Això protegeix les dades contra una possible corrupció quan es realitza la transferència.
- **MPI I/O:** MPI I/O són rutines MPI que permeten fer operacions de lectura i escriptura de fitxers en paral·lel. Lustre disposa d'una capa MPI ADIO que optimitza les I/O (in/out) paral·leles.
- **Exportació NFS i CIFS:** Els fitxers Lustre es poden exportar mitjançant NFS o CIFS. NFS (Network File System) és un protocol utilitzat per a sistemes d'arxius distribuït en un entorn de xarxa d'ordinadors locals. CIFS és un protocol propietat de Windows però que funciona perfectament amb Linux i MacOs. Gràcies al fet que Lustre disposa dels dos protocols es poden compartir arxius amb clients que no siguin Linux, com Windows o Mac Os X.
- **Eina de recuperació d'errors:** Lustre ofereix una eina que per restaurar la coherència entre els components d'emmagatzematge en cas que es produís un error greu en el sistema de fitxers. És destacable que un sistema de fitxers Lustre pot funcionar encara que hi hagi incoherències i que LFSSCK (Lustre File System Checker) es pot executar mentre el sistema de fitxers s'utilitza.
- **Supervisió de rendiment:** Lustre ofereix diversos mecanismes per examinar el rendiment del sistema.
- **Open source:** El programari Lustre està autoritzat sota la llicència GPL 2.0 per utilitzar-lo amb el sistema operatiu Linux.

Requeriments per utilitzar el sistema d'arxius Lustre

- Força potència de CPU, molta memòria i emmagatzematge ràpid al disc.
- Una bona amplada de banda del bus. A més és recomanable que l'emmagatzematge es distribueixi de forma uniforme entre els OSS.
- Respecte als clients han de disposar també d'un bon amplada de banda amb baixa latència.

4. Elecció del software

Després d'analitzar els diferents softwares vam optar per crear el nostre Parallel File System amb el software de Lustre. Les raons són senzilles, i és que vam trobar més informació en el referent a aquest software i això ens va permetre entendre millor el seu funcionament, cosa que ens ha ajudat a desenvolupar la part pràctica del projecte d'una forma una mica més ordenada i clara.

Els altres dos softwares també ens semblaven bones opcions però vam descartar GPFS perquè no és un software lliure, sinó de pagament, i BeeGFS per què no vam trobar tanta informació com el software Lustre.

A continuació farem una explicació més detallada dels components de Lustre, la seva xarxa i el seu funcionament.

4.1 Components Lustre

MGS

Lustre inclou un servidor de gestió (MGS) que s'encarrega d'emmagatzemar informació de configuració per als sistemes de fitxers Lustre que hi ha en un clúster i dona aquesta informació a altres components de Lustre.

Lustre contacta amb el MGS per proporcionar informació i els clients contacten amb el MGS per recuperar informació.

El més recomanable és que els MGS tinguin un emmagatzematge propi per poder-los gestionar de forma independent.

MDS

El MDS (Servidor de metadades) permet que les metadades emmagatzemades en un o més MDT estiguin disponibles pels clients Lustre. Cadascun dels MDS administra els sistemes d'arxius Lustre i administra també sol·licituds de xarxa per un o més MDT.

MDT

El MDT (Metadata Targets): emmagatzema metadades, com per exemple noms d'arxius, directoris, permisos, etc. És un emmagatzematge adjunt als MDS.

Cada sistema d'arxius disposa d'un MDT, i hem de tenir en compte que un MDT en un emmagatzematge compartit pot estar disponible per diversos MDS, tot i que només un MDS pot accedir a un MDT a la vegada. Amb això s'aconsegueix que si un MDS falla, un altre MDS pot accedir al MDT i posar les seves dades a disposició dels clients. Això es coneix com a commutació per error MDS.

OSS

L'OSS (Servidors d'Emmagatzematge d'Objectes) dona servei I/O (Entrada/Sortida) d'arxius i administra les sol·licituds de xarxa pels OST. Un OSS dona servei a entre dos i buit OST.

OST

L'OST (Objectiu d'emmagatzematge d'objectes), guarda els objectes que contenen informació de l'arxiu sobre l'usuari. Cada objecte es guarda en un OST que es troba en el sistema d'arxius Lustre.

Clients Lustre

Es consideren clients Lustre aquells nodes informàtics que executen el programari de client Lustre, cosa que els permet muntar un sistema d'arxius Lustre.

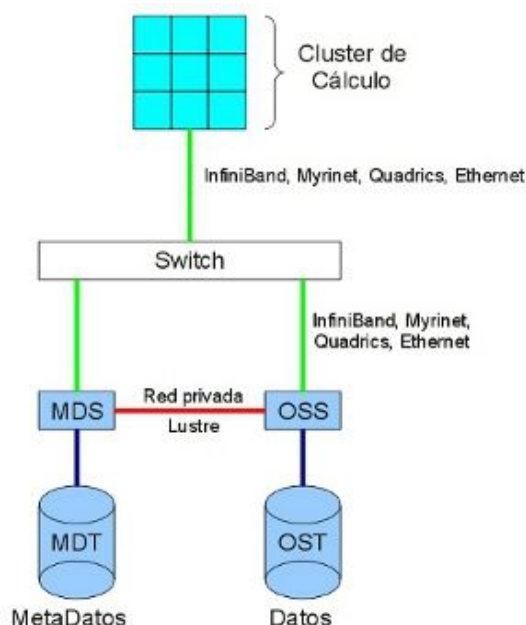
El programari de client de Lustre ofereix una interfície entre el sistema virtual d'arxius de Linux i els servidors Lustre. El programari de client inclou un client d'administració (MGC), un client de metadades (MDC) i clients d'emmagatzematge d'objectes múltiples (OSC), un corresponent a cada OST en el sistema d'arxius.

4.2 Xarxes Lustre

Una xarxa Lustre està formada per clients i servidors que executen el software Lustre. Lustre pot incloure diverses xarxes sempre que aquestes es puguin encaminar entre si. De forma semblant una sola xarxa pot tenir diverses subxarxes LNet. La xarxa Lustre està formada per dues capes, LNet i LND. La capa LNet opera per sobre de la capa LND de forma semblant en què la capa de xarxa opera per sobre de la capa d'enllaç de dades. La capa LNet no té connexió i no verifica les dades mentre que la capa LND està orientada a la connexió i per norma general verifica la transmissió de les dades.

4.3 Esquema del funcionament de Lustre

A continuació podem observar un molt aclaridor esquema que ens permet entendre molt millor el funcionament del sistema d'arxius Lustre.



Esquema funcionament de Lustre

En aquest esquema observem els components que hem esmentat i explicat anteriorment. A continuació veurem una petita explicació dels components basant-nos en l'esquema. Tenim:

- Un clúster, que seria el client de Lustre. Cada node computacional seria un client. Hem de destacar que cada node del clúster ha de portar instal·lat un agent de software amb suport per Lustre, ja que Lustre no utilitza els típics protocols per comunicar-se, com FTP o HTTP, sinó que utilitza un protocol propi.
- Interconnexions entre el clúster i els servidors. En l'esquema observem diverses connexions, tot i que una molt bona opció seria la InfiniBand, per la seva rapidesa. Tot i que les interconnexions es poden barrejar, és a dir, en una part del sistema podem tenir InfiniBand i en una altra part Ethernet, això no és l'ideal ni el més recomanable. Si totes les interconnexions són iguals això ajudarà al fet que el rendiment sigui constant.
- Una xarxa privada Lustre que és utilitzada pels servidors MDS i OSS per comunicar-se entre si.
- Un servidor de metadades (MDS), amb el seu emmagatzematge (un sistema de fitxers dedicat anomenat MDT). Aquest servidor com hem comentat s'encarrega de les metadades (permisos dels fitxers, dates de modificació, etc.).
- Un servidor de dades (OSS), amb el seu emmagatzematge (un sistema de fitxers dedicat anomenat OST). Aquest servidor s'encarrega d'emmagatzemar les dades.

5. Amazon Web Services (AWS)

Amazon Web Services (AWS) és una plataforma de serveis orientada al núvol a nivell global amb una gran varietat, tant com serveis d'infraestructura, de base de dades, d'emmagatzematge, de còmput i molts més.

Aquesta plataforma permet configurar les característiques d'aquests serveis que nosaltres utilitzarem, els serveis es paguen individualment i durant el temps que els utilitzem, és a dir, el cost va per hores utilitzades i depenen de les nostres característiques serà més car o més barat.

Nosaltres utilitzarem la instància Amazon Elastic Compute Cloud (Amazon EC2), és un servei que proporciona capacitat informàtica en el núvol de manera segura i modificable, molt similar a una màquina virtual però que en aquest cas es troba el núvol d'internet.

Abans d'executar el nostre servei, haurem de configurar-lo perquè el servei detecti altres instàncies en la seva xarxa, i també que nosaltres ens podem connectar des de casa aquest núvol i poder accedir aquesta xarxa per fer les modificacions adients perquè funcioni correctament el servei.

5.1 Configuració de la xarxa en AWS

En aquest apartat explicarem la preconfiguració de la xarxa, que és la que tindran totes les instàncies creades per nosaltres.

VPC:

Abans de procedir a la creació de la instància primer haurem de crear certes coses com la VPC (Virtual Private Cloud), la qual li assignarem el rang d'IPs que volem en aquest núvol, en el nostre cas la xarxa serà la 192.168.10.0/24.

[VPCs](#) > [Create VPC](#)

Create VPC

A VPC is an isolated portion of the AWS cloud populated by AWS objects, such as Amazon EC2 instances. You must specify an IPv4 a 10.0.0.0/16. You cannot specify an IPv4 CIDR block larger than /16. You can optionally associate an IPv6 CIDR block with the VPC.

Name tag	Lustre	
IPv4 CIDR block*	192.168.10.0/24	
IPv6 CIDR block	☒ No IPv6 CIDR Block ☐ Amazon provided IPv6 CIDR block ☐ IPv6 CIDR owned by me	
Tenancy	Default	

Subnet:

Creada la VPC haurem d'assignar-li la subnet que aquesta és la que s'encarregarà d'assignar-los la IP a les màquines si ho tenen amb automàtic, dintre del rang indicat entre la VPC i la Subnet.

[Subnets](#) > Create subnet

Create subnet

Specify your subnet's IP address block in CIDR format; for example, 10.0.0.0/24. IPv4 block sizes must be between a /16 netmask and /28 netmask.

Name tag ⓘ

VPC* ⓘ

Availability Zone ⓘ

VPC CIDRs	CIDR	Status
	192.168.10.0/24	associated

IPv4 CIDR block* ⓘ

Route Table:

A continuació s'haurà de crear una taula de ruta el qual especifica com s'envien aquests paquets entre les subnets del nostre VPC. Com que només tenim una subnet no hi haurà gaire problema però igualment s'ha d'introduir.

[Route Tables](#) > Create route table

Create route table

A route table specifies how packets are forwarded between the subnets within your VPC, the internet, and your VPN connection.

Name tag ⓘ

VPC* ⓘ

Internet Gateway:

I finalment la Gateway d'internet la qual serveix per connectar la VPC d'Amazon amb Internet, i d'aquesta manera poder connectar-nos aquest núvol des de qualsevol lloc, en el nostre cas des de casa.

[Internet gateways](#) > Create internet gateway

Create internet gateway

An internet gateway is a virtual router that connects a VPC to the internet. To create a new internet gateway specify the name for the gateway below.

Name tag ⓘ

5.2 Creació de la instància en AWS

Procedim a l'hora d'engegar la nostra instància d'AWS en el núvol, primer de tot haurem de seleccionar un sistema operatiu, en el nostre cas utilitzarem un Centos 7.7, concretament aquest sistema s'ha de buscar en el Marketplace perquè el sistema operatiu no pertany al free tier d'AWS.

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group

Step 1: Choose an Amazon Machine Image (AMI)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch an instance.

Search: centos

Quick Start (0)

My AMIs (0)

AWS Marketplace (321)

Community AMIs (4598)

Categories

All Categories

Infrastructure Software (265)

DevOps (127)

Business Applications (30)

CentOS 7 (x86_64) - with Updates HVM

★★★★★ (61) | 2002_01 | [Previous versions](#) | [By CentOS.org](#)

Linux/Unix, CentOS 7 | 64-bit (x86) Amazon Machine Image (AMI) | Updated: 3/16/20

Free tier eligible

This is the Official CentOS 7 x86_64 HVM image that has been built with a minimal set of packages.

[More info](#)

CentOS 6 (x86_64) - with Updates HVM

★★★★★ (34) | 2002_01 | [Previous versions](#) | [By CentOS.org](#)

Linux/Unix, CentOS 6 | 64-bit (x86) Amazon Machine Image (AMI) | Updated: 3/16/20

Free tier eligible

This is the Official CentOS 6 x86_64 HVM image that has been built with a minimal set of packages.

[More info](#)

Escollim el tipus d'instància que proveeix l'Amazon Machine Image (AMI) d'EC2, en el nostre cas hem escollit el tipus t2.micro el qual és gratuïta durant un mes en AWS.

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group

Step 2: Choose an Instance Type

Amazon EC2 provides a wide selection of instance types optimized to fit different use cases. Instances are virtual servers that can be launched in the cloud. Choose the appropriate mix of resources for your applications. [Learn more](#) about instance types and how they can meet your computing needs.

Filter by: **All instance types** **Current generation** [Show/Hide Columns](#)

Currently selected: t2.micro (Variable ECUs, 1 vCPUs, 2.5 GHz, Intel Xeon Family, 1 GiB memory, EBS only)

Note: The vendor recommends using a t2.micro instance (or larger) for the best experience with this product.

	Family	Type	vCPUs	Memory (GiB)
☐	General purpose	t2.nano	1	0.5
☒	General purpose	t2.micro Free tier eligible	1	1
☐	General purpose	t2.small	1	2

A continuació haurem de configurar els detalls de la nostra instància, tant la VPC (Virtual Private Cloud), com la subnet i assignar-li la IP que desitgem a la instància en aquell moment.

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign

Number of instances	1	Launch into Auto Scaling Group
Purchasing option	☐ Request Spot instances	
Network	vpc-02f6f02650aab1bb8 Lustre	Create new VPC
Subnet	subnet-0195c51237a6f1960 Lustre us-east-1c 244 IP Addresses available	Create new subnet
Auto-assign Public IP	Use subnet setting (Enable)	
Placement group	☐ Add instance to placement group	
Capacity Reservation	Open	Create new Capacity Reservation
IAM role	None	Create new IAM role
Shutdown behavior	Stop	
Stop - Hibernate behavior	☐ Enable hibernation as an additional stop behavior	
Enable termination protection	☐ Protect against accidental termination	
Monitoring	☐ Enable CloudWatch detailed monitoring Additional charges apply.	
Tenancy	Shared - Run a shared hardware instance Additional charges may apply when launching Dedicated instances.	
Elastic Inference	☐ Add an Elastic Inference accelerator Additional charges apply.	
T2/T3 Unlimited	☐ Enable Additional charges may apply	

Una vegada configurat els detalls de la instància, si volem podem afegir 1 o diversos discos a la instància, en el nostre cas només afegirem un. Si volem més rendiment podem escollir-lo en l'opció "Volume Type" i agafar un amb més IOPS.

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Add Tags 6. Configure Security Group 7. Review

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. [Learn more](#) about storage options in Amazon EC2.

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS
Root	/dev/sda1	snap-0cb4f5f601f70d3c	8	General Purpose SSD (gp2)	100 / 3000
EBS	/dev/sdb	Search (case-insensit	8	General Purpose SSD (gp2)	100 / 3000

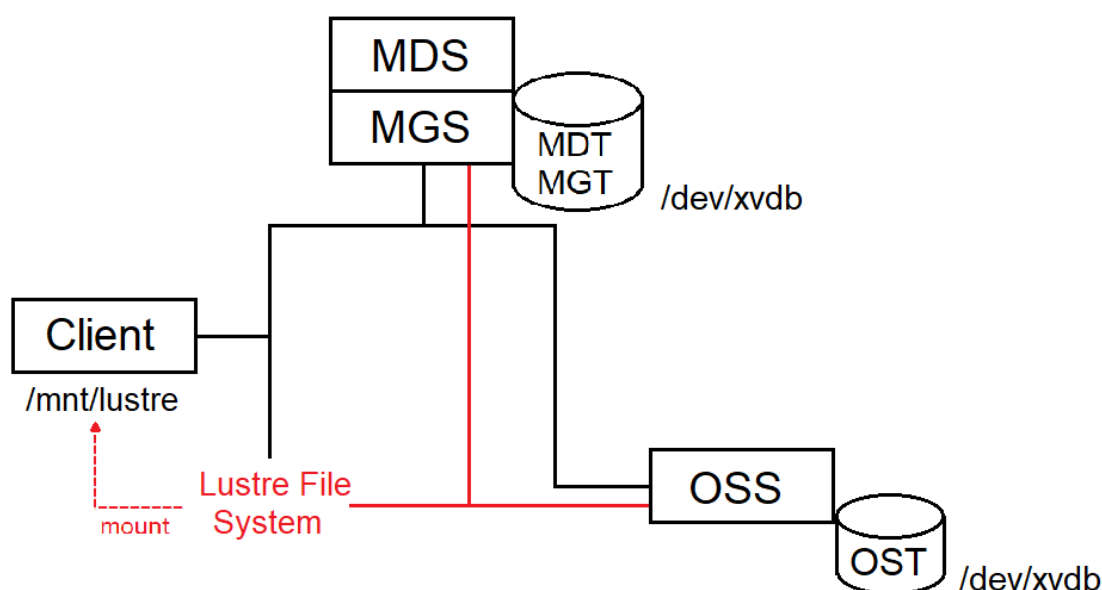
Add New Volume

6. Instal·lació de Lustre en CentOS 7 amb AWS

Un cop explicat el que és un Parallel File System ja podem començar amb la instal·lació d'un, en aquest cas un Parallel File System basat en el programari Lustre, que és el que hem escollit com ja hem explicat anteriorment.

Per crear el nostre Parallel File System utilitzarem AWS (Amazon Web Service). Aquesta eina ens permet crear màquines virtuals amb diferents característiques de forma ràpida i eficient. A més aquestes màquines s'emmagatzemen i corren en servidors externs, cosa que ens permet treballar de forma fluida sense utilitzar els recursos del nostre ordinador.

El nostre Parallel File System estarà muntat en tres nodes diferents (és a dir tres màquines) amb un SO CentOS 7.7. A continuació es mostra un esquema de com és el nostre sistema d'arxius i com estan distribuïts els nodes.



Esquema del nostre sistema Lustre

Com ja hem explicat el sistema d'arxius Lustre està compost per tres tipus de servidors: MGS (Servidor de gestió), MDS (Servidor de metadades) i OSS (Servidor d'emmagatzematge d'objectes).

Primer observem en l'esquema anterior els servidors MGS i MDS. Aquests dos servidors estaran creats en un mateix node. La informació dels servidors MGS i MDS s'emmagatzema al MGT (Management Target) i MDT (Management Target), que estaran al disk /dev/xvdb.

Després observem l'OSS, que és un servidor que emmagatzema les dades per tot el contingut dels arxius i que es trobarà instal·lat en el segon node. Aquestes dades són emmagatzemades per l'OST (Object Storage Targets), que es troba al disk /dev/xvdb.

El Client munta el sistema de fitxers Lustre en el directori /mnt/lustre i el fa servir com un directori muntat NFS que té capacitats d'alt rendiment. En un supercomputador, generalment hi ha molts nodes informàtics: els amfitrions a la xarxa interna d'un supercomputador es denominen "nodes", i aquests nodes es converteixen en clients de Lustre.

6.1 Creació part comuna nodes

Per començar amb la instal·lació haurem d'agregar els tres nodes a l'arxiu /etc/hosts de tots els nodes. L'arxiu /etc/hosts es veuria de la següent manera:

127.0.0.1 localhost

192.168.10.120 client

192.168.10.50 mgs

192.168.10.102 oss

NOTA: Això ho hem de fer en tots els nodes, també al client.

A continuació, instal·larem els repositoris EPEL, ZFS, i Chrony. A més també configurarem la data i hora dels nodes perquè tot funcioni correctament:

```
$ sudo yum -y install epel-release
$ rpm -ivh http://download.zfsonlinux.org/epel/zfs-release.el7.centos.noarch.rpm
$ sudo yum -y install chrony
$ sudo systemctl start chronyd
$ sudo systemctl enable chronyd
$ sudo timedatectl set-timezone Europe/Madrid
```

6.2 Instal·lació dels repositoris dels servidors

Els següents repositoris hauran de ser instal·lats només als nodes que actuaran com a servidors. Tant al node que combina MGS i MDS com al node amb OSS.

Creem l'arxiu de configuració del repositori Lustre a /etc/yum.repos.d/lustre.repo i fiquem el següent:

```
[lustre-server]
name=lustre-server
baseurl=https://downloads.whamcloud.com/public/lustre/latest-release/el7/server
# exclude=debuginfo
gpgcheck=0

[lustre-client]
name=lustre-client
baseurl=https://downloads.whamcloud.com/public/lustre/latest-release/el7/client
# exclude=debuginfo
gpgcheck=0

[e2fsprogs-wc]
name=e2fsprogs-wc
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/el7
# exclude=debuginfo
gpgcheck=0
```

Instal·lem els següents repositoris necessaris perquè els servidors Lustre funcionin correctament:

```
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/kernel-3.10.0-1062.1.1.el7\_lustre.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/kmod-lustre-2.13.0-1.el7.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/lustre-osd-ldiskfs-mount-2.13.0-1.el7.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/kernel-headers-3.10.0-1062.1.1.el7\_lustre.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/kernel-devel-3.10.0-1062.1.1.el7\_lustre.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/kmod-lustre-osd-ldiskfs-2.13.0-1.el7.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/lustre/latest-feature-release/el7/server/RPMS/x86\_64/lustre-2.13.0-1.el7.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/e2fsprogs/1.45.2.wc1/el7/RPMS/x86\_64/e2fsprogs-1.45.2.wc1-0.el7.x86\_64.rpm
$ sudo yum -y install
https://downloads.whamcloud.com/public/e2fsprogs/1.45.2.wc1/el7/RPMS/x86\_64/e2fsprogs-libs-1.45.2.wc1-0.el7.x86\_64.rpm
```

Un cop instal·lats els repositoris haurem de reiniciar el node per arrencar-ho novament però aquest cop amb el nucli Lustre habilitat. A més haurem d'iniciar el servei Lustre amb la següent comanda:

```
$ systemctl start lustre
```

Perquè el servei Lustre s'iniciï cada cop que reiniciem el mòdul executarem la següent comanda:

```
$ systemctl enable lustre
```

Per veure els repositoris instal·lats i comprovar que estan tots els necessaris podem executar la següent comanda:

```
# Comprobación de paquetes instalados de lustre:  
$ rpm -aq | grep -i lustre
```

Per esborrar un repositori instal·lat podem executar la següent comanda:

```
# Borrar un paquete instalado:  
$ rpm -e <nombre_paquete>
```

6.3 Mòdul LNET als servidors (MGS-MDS i OSS)

Creem l'arxiu a `/etc/modprobe.d/lnet.conf` per configurar una LNET amb el paràmetre de xarxa `tcp0 (eth0)`. Només necessitem afegir el següent a l'arxiu:

```
$ sudo vi /etc/modprobe.d/lnet.conf
> options lnet networks="tcp0(eth0)"
```

LNET és el protocol de comunicació de xarxa utilitzat per Lustre. L'opció que hem afegit, `networks=tcp0(eth0)`, indica que el node forma part de la LNET amb nom `tcp0` i que utilitza la interfície de xarxa `eth0`.

A continuació carregarem el mòdul LNET de forma manual amb la següent comanda:

```
$ sudo modprobe lnet
$ lnetctl lnet configure
```

Comprovem que LNET està carregat correctament:

```
$ sudo lsmod | grep lnet
```

Afegir el node en LNET de la xarxa `tcp0`, en la interfície `eth0`:

```
$ sudo lnetctl net add --net tcp0 --if eth0
# Mostrar la configuració de LNET
$ sudo lnetctl net show
```

Para carregar automàticament els mòduls LNet en l'arranc, hem de crear l'arxiu `/etc/sysconfig/modules/lnet.modules` i inserir el següent script en l'arxiu:

```
$ sudo vi /etc/sysconfig/modules/lnet.modules
#!/bin/sh
if [ ! -c /dev/lnet ] ; then
    exec /sbin/modprobe lnet >/dev/null 2>&1
fi
```

6.4 Instal·lació dels repositoris del Client

Els següents repositoris hauran de ser instal·lats únicament als nodes que actuïn com a clients. La instal·lació es farà a partir dels repositoris AWS corresponents per la versió kernel del nostre SO CentOS 7.7.

Els passos a seguir per afegir els repositoris i instal·lar-los són els següents

```
$ sudo yum -y install wget
$ sudo wget https://fsx-lustre-client-repo-public-keys.s3.amazonaws.com/fsx-rpm-public-key.asc -O /tmp/fsx-rpm-public-key.asc
$ sudo rpm --import /tmp/fsx-rpm-public-key.asc
$ sudo wget https://fsx-lustre-client-repo.s3.amazonaws.com/el/7/fsx-lustre-client.repo -O /etc/yum.repos.d/aws-fsx.repo
$ cat /etc/centos-release
$ sudo sed -i 's#7#7.7#' /etc/yum.repos.d/aws-fsx.repo
$ sudo yum clean all
$ sudo yum install -y kmod-lustre-client lustre-client
```

Comandes útils en la instal·lació:

```
# Comprovació de paquets instal·lats de lustre:
$ rpm -aq | grep -i lustre
# Borrar un paquete instalado:
$ rpm -e <nombre_paquete>
```

6.5 Activació del mòdul Lustre al Client

Un cop instal·lats els repositoris corresponents haurem d'habilitar el mòdul Lustre al Client per poder muntar el nostre sistema de fitxers. Per fer-ho executem la següent comanda:

```
$ sudo modprobe lustre
```

Per comprovar si el mòdul Lustre està correctament carregat executem la següent comanda:

```
$ sudo lsmod | grep lustre
```

Per carregar automàticament el mòdul Lustre en iniciar el sistema, hem de crear el fitxer `/etc/sysconfig/modules/lustre.modules` amb l'script següent:

```
#!/bin/sh
/sbin/lsmod | /bin/grep lustre 1>/dev/null 2>&1
if [ ! $? ] ; then
    /sbin/modprobe lustre >/dev/null 2>&1
fi
```

6.6 Muntatge del Lustre File System

Un cop realitzats els passos anteriors ja estem preparats per muntar el nostre sistema d'arxius Lustre. Per fer-ho haurem de seguir els següents passos depenent del mòdul en el qual estiguem.

Crear MGS & MDS:

En aquest apartat mostrarem les comandes necessàries per habilitar el nostre node MGS-MDS:

```
# Formatar el disc xvdb en format Lustre i preparar-ho per funcionar com a node MGS-MDS
$ sudo mkfs.lustre --fsname=DATA --mgs --mdt --index=0 /dev/xvdb
# Creem el directori DATA que és on muntarem el disc que acabem de formatar, el xvdb
$ sudo mkdir /DATA
# Muntem el disc xvdb al directori /DATA
$ sudo mount -t lustre /dev/xvdb /DATA
# Executem blkid per veure el UUID del disc xvdb i afegir una entrada a l'arxiu /etc/fstab perquè el disc es munti automàticament en reiniciar el node.
$ sudo blkid
# Afegim l'entrada al /etc/fstab
$ sudo vi /etc/fstab
> UUID=032 /DATA lustre defaults 0 0
```

Crear OST en OSS:

En aquest apartat mostrarem les comandes necessàries per habilitar el nostre node OST:

```
# Formatar el disc xvdb en format Lustre i preparar-ho per funcionar com a node OST
$ sudo mkfs.lustre --fsname=DATA --ost --mgsnode=mgs@tcp0 --index=0 /dev/xvdb
# Creació del directori ost, on es muntarà el disc que acabem de formatar, el xvdb
$ sudo mkdir /ost
# Muntem el disc xvdb al directori /ost
$ sudo mount -t lustre /dev/xvdb /ost
# Executem blkid per veure el UUID del disc xvdb i afegir una entrada a l'arxiu /etc/fstab perquè el disc es munti automàticament en reiniciar el node.
$ sudo blkid
# Afegim l'entrada al /etc/fstab
$ sudo vi /etc/fstab
> UUID=032 /ost lustre defaults 0 0
```

Muntar Lustre file system en el Client:

En aquest apartat mostrarem les comandes necessàries per muntar el nostre sistema d'arxius en el node client:

```
#Creem la carpeta en la qual muntarem el nostre sistema d'arxius Lustre ja creat
$ sudo mkdir /mnt/lustre
# Muntem el sistema Lustre que es troba a mgs@tcp0:/DATA al directori anteriorment creat, /mnt/lustre
$ sudo mount -t lustre mgs@tcp0:/DATA /mnt/lustre
```

6.7 Comprovació del Sistema de fitxers Lustre

Un cop instal·lat el nostre sistema de fitxers és el moment de comprovar si funciona correctament. Per fer-ho haurem d'anar al node client i fer comprovacions amb la comanda **lfs**. A continuació mostrarem el resultat de les proves que nosaltres hem realitzat al nostre sistema de fitxers.

Check del punt de muntatge /mnt/lustre:

```
$ sudo df -h /mnt/lustre

[centos@Client1 ~]$ sudo df -h /mnt/lustre
S. fitxers          Mida En ús Lliure  %Ús Muntat a
192.168.10.200@tcp:/DATA  7,4G   38M   7,0G    1% /mnt/lustre
[centos@Client1 ~]$
```

Check del servidor:

```
$ sudo lfs check servers

[centos@Client1 ~]$ sudo lfs check servers
DATA-OST0000-osc-ffff97abf9e08000 active.
DATA-MDT0000-mdc-ffff97abf9e08000 active.
[centos@Client1 ~]$
```

Check the file system:

```
$ sudo lfs df -h

[centos@Client1 ~]$ sudo lfs df -h
UUID          bytes      Used    Available Use% Mounted on
DATA-MDT0000_UUID  4.4G    33.9M      4.0G    1% /mnt/lustre[MDT:0]
DATA-OST0000_UUID  7.4G    37.1M      6.9G    1% /mnt/lustre[OST:0]

filesystem_summary:    7.4G    37.1M      6.9G    1% /mnt/lustre
[centos@Client1 ~]$
```

7. Script d'instal·lació de Lustre

Per aixecar el nostre servei de forma ràpida, senzilla i en qualsevol infraestructura amb un sistema operatiu CentOS 7.7 hem decidit crear un script a partir del qual es pot aixecar tot el sistema d'arxius Lustre. En aquest script hi ha 3 opcions. La instal·lació del MGS-MDS, la instal·lació de l'OST i la del Client. A continuació adjuntarem dos enllaços, un on es pot trobar l'script creat per nosaltres i el qual utilitzem al vídeo, i un altre vídeo de demostració de com funciona el script i com a partir d'aquest podem aixecar el Parallel File System Lustre.

- Link de l'script:

<https://github.com/FinalProjectLustre/LustreInstall/blob/master/Script>

- Link del video aixecant Lustre a partir del nostre script:

https://drive.google.com/file/d/1JP__ja-vKE1r1ifcTvJCRT9cf43riXXv/view?usp=sharing

8. Conclusió

En conclusió podríem dir que el projecte ens ha donat nous coneixements i ens ha permès conèixer una altra branca de la informàtica que estar en creixement exponencial en l'actualitat. Quan vam començar aquest projecte no sabíem gairebé res en el referent a la supercomputació i els Parallel File System, però un cop acabat podem dir que hem après molt del tema relacionat.

Com hem dit, en començar el projecte no teníem coneixements sobre el tema i això ens va dificultar les coses al principi, ja que no enteníem gairebé res. Per resoldre aquest problema vam fer una investigació especialment profunda sobre el tema, i no vam començar la part pràctica fins no tenir una base clara sobre els conceptes bàsics d'un Parallel File System.

Un cop començada la part pràctica, és a dir, la creació del nostre propi Parallel File System amb Lustre, també vam trobar bastants dificultats. No trobàvem manuals totalment fiables i que funcionessin correctament, per aquest motiu vam anar utilitzant diversos manuals recolzant-nos sempre en el manual oficial de Lustre que vam trobar. A més també vam rebre un molt útil assessorament per part d'alguns integrants de l'empresa HPCNow. Vam dedicar moltes hores i proves fins a aconseguir els resultats desitjats, però els errors que anàvem cometent ens van ajudar a aprendre i a entendre cada cop més com funciona Lustre.

Un cop acabada la instal·lació del nostre sistema Lustre i comprovar que funcionava correctament ens va semblar que això no era suficient i vam enfocar el projecte a un nivell més, crear un script a partir del qual poder muntar tot el sistema d'arxius Lustre.

Respecte a les coses que ens haurien agradat poder fer, destacaríem el fet d'haver muntat tot el sistema en un entorn més real, basat per exemple en diverses raspberrys, però donada la situació no ha estat possible. Tot i això no tot ha estat dolent en el relacionat amb això, ja que hem après com funciona Amazon Web Service (AWS), i ens ha semblat una eina increïblement útil.

Per finalitzar aquesta conclusió podríem dir que ens alegra haver realitzat aquest projecte, pels coneixements adquirits, pel repte que ha suposat i per les increïbles noves tecnologies que hem pogut utilitzar i sobre les quals no sabíem res.

9. Glossari

- **Metadades:** És la via per a comunicar informació sobre un document o sobre els recursos que directament es relacionen amb la seva accessibilitat, serveixen per identificar, autenticar i contextualitzar arxius.
- **Redundància:** És la duplicació de les dades o del maquinari de caràcter crític que es vulgui assegurar davant dels possibles errors que puguin sorgir durant un ús continuat.
- **POSIX:** Acrònim de Portable Operating System Interface, i X ve d'UNIX com a senyal d'identitat de l'API.
- **Striping:** Cada arxiu és distribuït en parts d'una determinada mida i aquestes parts són distribuïdes pels servidors d'emmagatzematge existents.
- **ZFS:** Sistema d'arxius i administrador de volums desenvolupat originalment per Sun Microsystems per al seu sistema operatiu Solaris. El seu acrònim significa 'Zettabyte File System'.
- **RDMA:** Remote direct memory access o la seva traducció a català, accés remot directe a memòria consisteix en l'accés directe des de la memòria principal d'un ordinador a la d'un altre sense cooperació de sistema operatiu.
- **InfiniBand:** Bus de comunicacions sèrie d'alta velocitat, baixa latència i de baixa sobrecàrrega de CPU, dissenyat tant per a connexions internes com externes.
- **LDLM:** Lustre Distributed Lock Manager és la infraestructura de bloqueig de Lustre que gestiona els bloquejos entre clients i servidors i els bloquejos locals a un node.
- **I/O:** El terme I / O s'utilitza per a descriure qualsevol programa o transmetre dades des d'un sistema i cap a un dispositiu perifèric o viceversa.
- **LNet:** Lightweight Networking Library és una col·lecció de classes i components per permetre la creació de xarxes TCP o UDP.
- **Amazon Web Services:**
 - **AMI:** Una imatge d'una màquina d'Amazon, similar a la VDI de VirtualBox, es una plantilla la qual predeterminadament conté una configuració de software.
 - **Instància:** És un servidor virtual en el cloud, cadascuna té les seves característiques de processadors, memòria, xarxa i emmagatzematge.

10. Webgrafia

- **Parallel File System:**
 - https://en.wikipedia.org/wiki/Clustered_file_system
 - <https://www.linkedin.com/pulse/parallel-file-system-vs-network-dummies-briti-gangopadhay>
 - <https://arxiv.org/ftp/arxiv/papers/1904/1904.03997.pdf> (Lustre/GPFS)
 - http://www.idris.fr/media/docs/docu/idris/idris_patc_filesystems_proj.pdf
 - https://web.archive.org/web/20131224101831/http://www.snia-europe.org/objects_store/Christian_Bandulet_SNIATutorial%20Basics_EvolutionFileSystems.pdf
 - <http://www.cs.iit.edu/~iraicu/teaching/CS554-F13/lecture17-pfs-sam-lang.pdf>
- **Tipus de PFS:**
 - <http://etutorials.org/Linux+systems/cluster+computing+with+linux/Part+III+Managing+Clusters/Chapter+19+Parallel+I+O+and+the+Parallel+Virtual+File+System/19.2+Parallel+File+System+Architectures/>
- **IOPS:**
 - <https://en.wikipedia.org/wiki/IOPS>
 - <https://www.techrepublic.com/blog/the-enterprise-cloud/calculate-iops-in-a-storage-array/>
- **IPS:**
 - https://en.wikipedia.org/wiki/Instructions_per_second#Timeline_of_instructions_per_second
- **BeeGFS:**
 - <https://www.beegfs.io/wiki/TableOfContents>
 - <https://en.wikipedia.org/wiki/BeeGFS>
 - https://www.beegfs.io/wiki/FAQ#hn_59ca4f8bbb_1
- **GPFS IBM**
 - https://es.wikipedia.org/wiki/General_Parallel_File_System
 - https://www.ibm.com/support/knowledgecenter/en/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.product.doc/doc/bi_gpfs_overview.html
- **Lustre:**
 - <https://arxiv.org/ftp/arxiv/papers/1904/1904.03997.pdf>
 - http://doc.lustre.org/lustre_manual.xhtml#understandinglustre.whatislustre
 - <https://www.muylinux.com/2010/02/17/oracle-mantiene-lustre/>
 - **MGS:**
 - [http://wiki.lustre.org/Lustre_Management_Service_\(MGS\)](http://wiki.lustre.org/Lustre_Management_Service_(MGS))
 - **MDS:**
 - [http://wiki.lustre.org/Lustre_Metadata_Service_\(MDS\)](http://wiki.lustre.org/Lustre_Metadata_Service_(MDS))
 - **OSS:**
 - [http://wiki.lustre.org/Lustre_Object_Storage_Service_\(OSS\)](http://wiki.lustre.org/Lustre_Object_Storage_Service_(OSS))
- **Manual Lustre:**
 - http://doc.lustre.org/lustre_manual.pdf