Lustre File System

Lustre File System es un sistema paralelo de código abierto que está adaptado para numerosos requisitos de entornos de simulación HCP (Dynamic Host Configuration).

Proporciona una interfaz de sistema de archivos compatible con el estándar POSIX, atender a miles de clientes y gestionar petabytes de almacenamiento con un ancho de banda de cientos de GBps.

Un sistema de archivos Lustre tiene tres grandes unidades funcionales:

- Uno o más servidores de metadatos (MDS) nodos que tiene una o más metadatos de destino (MDT) dispositivos por Lustre sistema de archivos que almacena metadatos espacio de nombres, tales como nombres de archivos, directorios, permisos de acceso y disposición de los archivos.

Los datos MDT se almacenan en un sistema de archivos del disco local. El servidor de metadatos Lustre sólo está implicado en nombre de ruta y la gestión de permisos, en ningún caso en los ficheros de E/S, favoreciendo la escalabilidad en el servidor de metadatos. La capacidad de tener múltiples EMD en un único sistema de archivos es una característica de Lustre 2.4, y permite a los subárboles de directorio residen en los EMD secundarios, mientras que en las versiones de Lustre 2.7 en adelante también grandes directorios individuales pueden ser distribuidos a través de múltiples EMD.

- Uno o más servidores de almacenamiento de objetos (OSS) nodos que almacenan datos de archivos en uno o más de destino de almacenamiento de objetos (OST) dispositivos. Dependiendo del hardware del servidor, un OSS sirve normalmente entre dos y ocho OST, con cada OST gestión de un único sistema de archivos del disco local. La capacidad de un sistema de archivos Lustre es la suma de las capacidades proporcionadas por los OST.

- Cliente(s) de acceso y uso de los datos. Lustre presenta todos los clientes con un espacio de nombres unificado para todos los archivos y datos en el sistema de archivos, utilizando estándares POSIX la semántica, y permite simultáneas de lectura y coherente, y escribe el acceso a los archivos en el sistema de archivos.

El lustre de la red (LNET) capa puede utilizar varios tipos de interconexiones de red, incluyendo verbos nativos InfiniBand, Omni-Path , RoCE y iWARP vía OFED, TCP / IP sobre Ethernet y otras tecnologías de red propios, como el Cray interconexión Géminis. En Lustre 2.3 y anteriores, Myrinet , Quadrics , Cray SeaStar y redes RapidArray también fueron apoyados, pero estos controladores de red quedaron en desuso cuando estas redes ya no estaban disponibles comercialmente, y el soporte se eliminó completamente en Lustre 2,8. Lustre aprovechará remoto de acceso directo a memoria (Transferencias RDMA) cuando esté disponible, para mejorar el rendimiento y reducir el uso de CPU.

El almacenamiento utilizado para los sistemas de ficheros de respaldo MDT y OST se proporciona normalmente por los dispositivos RAID, aunque funciona con cualquier dispositivo de bloque. Desde Lustre 2.4, el MDT y OST también pueden utilizar ZFS para el sistema de ficheros de soporte además de ext4 , que les permite utilizar eficazmente JBOD de almacenamiento en lugar de los dispositivos RAID hardware.

Los servidores Lustre OSS y MDS leen, escriben y modifican datos en el formato impuesto por el sistema de archivos de respaldo y devuelven estos datos a los clientes. Esto permite a Lustre para tomar ventaja de las mejoras y características en el sistema de archivos subyacente, tal como compresión de datos y de las sumas de comprobación en ZFS. Los clientes no tienen acceso directo al almacenamiento subyacente, lo que asegura que un cliente torpe o malicioso no pueda corromper la estructura del sistema de archivos.

Un OST es un sistema de archivos dedicado que exporta una interfaz a niveles de bytes de los objetos de archivo para las operaciones de lectura / escritura. Un MDT es un sistema de archivos dedicado que almacena ínodos, directorios, POSIX y atributos de archivo ampliados , controla el archivo de permisos de acceso / ACL , y le dice a los clientes el diseño del objeto (s) que componen cada archivo normal. EMD y OST utilizan actualmente, ya sea una versión mejorada de ext4 llamada ldiskfs o ZFS / DMU para el almacenamiento de datos de back-end para almacenar archivos / objetos utilizando el código abierto ZFS-on-Linux puerto.

Cuando un cliente accede a un archivo, se realiza una búsqueda de nombre de archivo en el MDS. Cuando las operaciones de búsqueda de nombre de archivo MDS es completo y el usuario y el cliente tiene permiso para acceder y / o crear el archivo, ya sea el diseño de un archivo existente se devuelve al cliente o un nuevo archivo se crea en nombre del cliente, si así lo solicita. Para las operaciones de lectura o escritura, el cliente interpreta la disposición de los archivos en el volumen de objeto lógico (LOV) capa, que mapea el archivo lógico desplazamiento y el tamaño de uno o más objetos, cada uno que reside en un OST separada. Luego, el cliente cierra la gama de archivos que se opera y ejecuta una o más paralelos leer o escribir directamente a las operaciones de los nodos de OSS. Con este enfoque, los cuellos de botella para las comunicaciones cliente-a-OSS son eliminados, por lo que el ancho de banda total disponible para los clientes a leer y escribir datos de escalas casi linealmente con el número de OST en el sistema de archivos.

Tras la búsqueda inicial de la estructura del archivo, el MDS no participa normalmente en las operaciones de IO de archivos ya que toda asignación de bloque y datos IO se gestiona internamente por el OST. Los clientes no modifican directamente los objetos o datos en los sistemas de archivos OST, sino que delegan esta tarea a los nodos de OSS. Este enfoque garantiza la escalabilidad de las agrupaciones y superordenadores a gran escala, así como una mayor seguridad y fiabilidad. Por el contrario, los sistemas de archivos basados en bloques compartidos como GPFS y OCFS permiten el acceso directo al almacenamiento

subyacente por todos los clientes en el sistema de archivos, lo que requiere un gran back-end SAN unido a los clientes, y aumenta el riesgo de corrupción del sistema de archivos del mal comportamiento / clientes defectuosos.

Autor: Adrián Izquierdo Pozo