

DELL EMC POWERSCALE ONEFS: RESUMEN TÉCNICO

Resumen

En esta documentación técnica, se proporcionan detalles técnicos sobre las características y funcionalidades clave del sistema operativo OneFS que se utiliza para potenciar todas las soluciones de almacenamiento NAS de escalamiento horizontal de Dell EMC PowerScale.

Septiembre de 2021

Revisiones

Versión	Fecha	Comentario
1.0	Noviembre de 2013	Versión inicial de OneFS 7.1
2.0	Junio de 2014	Actualización para OneFS 7.1.1
3.0	Noviembre de 2014	Actualización para OneFS 7.2
4.0	Junio de 2015	Actualización para OneFS 7.2.1
5.0	Noviembre de 2015	Actualización para OneFS 8.0
6.0	Septiembre de 2016	Actualización para OneFS 8.0.1
7.0	Abril de 2017	Actualización para OneFS 8.1
8.0	Noviembre de 2017	Actualización para OneFS 8.1.1
9.0	Febrero de 2019	Actualización para OneFS 8.1.3
10.0	Abril de 2019	Actualización para OneFS 8.2
11.0	Agosto de 2019	Actualización para OneFS 8.2.1
12.0	Diciembre de 2019	Actualización para OneFS 8.2.2
13.0	2020 de junio	Actualización para OneFS 9.0
14.0	Septiembre de 2020	Actualización para OneFS 9.1
15.0	Abril de 2021	Actualización para OneFS 9.2
16.0	Septiembre de 2021	Actualización para OneFS 9.3

Agradecimientos

La producción de esta documentación estuvo a cargo de las siguientes personas:

Autor: Nick Trimbee

La información de esta publicación se proporciona "tal cual". Dell Inc. no se hace responsable ni ofrece garantía de ningún tipo con respecto a la información de esta publicación y desconoce específicamente toda garantía implícita de comerciabilidad o capacidad para un propósito determinado.

El uso, la copia y la distribución de cualquier software descrito en esta publicación requieren una licencia de software correspondiente.

Copyright © Dell Inc. o sus filiales. Todos los derechos reservados. Dell, EMC, Dell EMC y otras marcas comerciales son marcas comerciales de Dell Inc. o sus filiales. Las demás marcas comerciales pueden ser marcas comerciales de sus respectivos dueños.

TABLA DE CONTENIDO

Introducción	4
Visión general de OneFS.....	5
Nodos PowerScale.....	5
Red.....	6
Visión general del software OneFS	7
Estructura del sistema de archivos.....	11
Diseño de datos y.....	12
Escritura de archivos.....	13
Almacenamiento en caché de OneFS	16
Coherencia de caché de OneFS.....	18
Caché de nivel 1.....	19
Caché de nivel 2.....	19
Caché de nivel 3.....	19
Lecturas de archivos	21
Bloqueos y simultaneidad	22
I/O multiproceso	24
Protección de datos	24
Compatibilidad.....	32
Protocolos compatibles	33
Operaciones no disruptivas: compatibilidad con protocolos	34
Filtrado de archivo.....	34
Desduplicación de datos: SmartDedupe.....	34
Eficiencia del almacenamiento de archivos pequeños	35
Reducción de datos en línea.....	36
Interfaces.....	39
Autenticación y control de acceso.....	39
Active Directory	40
Zonas de acceso	40
Administración basada en funciones	41
Auditoría de OneFS.....	41
Actualización de software.....	41
Software de administración y protección de datos de OneFS.....	43
Conclusión	44
DÉ UN PASO ADELANTE.....	44

Introducción

Las tres capas del modelo de almacenamiento tradicional (sistema de archivos, administrador de volúmenes y protección de datos) evolucionaron con el tiempo para adaptarse a las necesidades de las arquitecturas de almacenamiento de escala pequeña, pero presentan una complejidad considerable y no están bien adaptadas a los sistemas de escala de petabytes. El sistema operativo OneFS reemplaza todas estas por medio de un sistema de archivos en clúster unificador con protección de datos escalable incorporada, y obvia la necesidad de administración de volúmenes. OneFS es un componente básico fundamental para las infraestructuras de escalamiento horizontal, lo que permite una escala masiva y una eficiencia increíble, y se utiliza para potenciar todas las soluciones de almacenamiento NAS de Dell EMC PowerScale.

De manera crucial, OneFS está diseñado para escalar no solo en términos de máquinas, sino también en términos humanos, lo que permite que los sistemas a gran escala se administren con una fracción del personal que se requiere para los sistemas de almacenamiento tradicionales. OneFS elimina la complejidad e incorpora la funcionalidad de autoadministración y autorreparación que reduce drásticamente la carga de la administración de almacenamiento. OneFS también incorpora paralelismo en un nivel muy profundo del sistema operativo, por lo que prácticamente todos los servicios clave del sistema se distribuyen entre varias unidades de hardware. Esto permite que OneFS escale en casi todas las dimensiones a medida que se expande la infraestructura, lo que garantiza que lo que funciona hoy en día continúe funcionando a medida que crece el conjunto de datos.

OneFS es un sistema de archivos totalmente simétrico sin punto único de falla, ya que aprovecha la agrupación en clústeres no solo para escalar el rendimiento y la capacidad, sino también para permitir la conmutación por error universal y varios niveles de redundancia que van más allá de las funcionalidades de RAID. La tendencia de los subsistemas de disco ha sido aumentar lentamente el rendimiento y aumentar con rapidez la densidad de almacenamiento. OneFS responde a esta realidad escalando la cantidad de redundancia, así como la velocidad de la reparación de fallas. Esto permite que OneFS crezca a una escala de múltiples petabytes y proporcione una mayor confiabilidad que los sistemas de almacenamiento pequeños y tradicionales.


El hardware de PowerScale proporciona el dispositivo en el que se ejecuta OneFS. Los componentes de hardware son los mejores en su clase, pero están basados en hardware genérico, lo que garantiza que el hardware aproveche las curvas de eficiencia y costo de hardware genérico que mejoran constantemente. OneFS permite que el hardware se incorpore o se elimine del clúster a voluntad y en cualquier momento, lo que abstrae los datos y las aplicaciones del hardware. La longevidad de los datos es infinita y está protegida por la vicisitudes de las generaciones de hardware en evolución. Se eliminan el costo y los inconvenientes de las migraciones de datos y las actualizaciones de hardware.

OneFS es ideal para aplicaciones basadas en archivos y no estructuradas de "Big Data" en entornos empresariales, que incluyen directorios principales a gran escala, recursos compartidos de archivos, archivos, virtualización y análisis comerciales. Por lo tanto, OneFS se utiliza ampliamente en muchos sectores con uso intensivo de datos en la actualidad, como los sectores de energía, servicios financieros, Internet y servicios de alojamiento, inteligencia comercial, ingeniería, fabricación, medios de comunicación y entretenimiento, bioinformática, investigación científica y otros entornos de computación de alto rendimiento.

Público objetivo

En este informe, se presenta información sobre la implementación y la administración de clústeres de Dell EMC PowerScale y se proporciona un contexto integral para la arquitectura de OneFS.

El público objetivo de esta documentación técnica es cualquier usuario que configure y administre un entorno de almacenamiento en clúster de PowerScale. Se da por sentado que el lector tiene una comprensión básica del almacenamiento, las redes, los sistemas operativos y la administración de datos.

 Encontrará más información sobre los comandos y la configuración de funciones de OneFS en la [Guía de administración de OneFS](#).

Visión general de OneFS

OneFS combina las tres capas de arquitecturas de almacenamiento tradicionales (sistema de archivos, administrador de volúmenes y protección de datos) en una sola capa de software unificada, con lo que se crea un único sistema de archivos distribuido e inteligente que se ejecuta en un clúster de almacenamiento con tecnología OneFS.

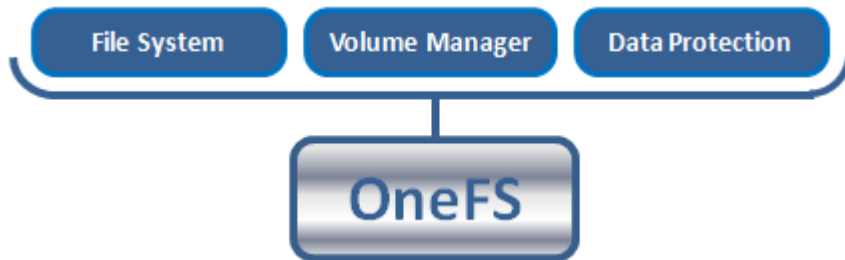


Figura 1: OneFS combina el sistema de archivos, el administrador de volúmenes y la protección de datos en un solo sistema distribuido e inteligente.

Esta es la innovación principal que permite directamente a las empresas utilizar de manera exitosa el NAS de escalamiento horizontal en sus entornos de hoy en día. Cumple con los principios clave del escalamiento horizontal: software inteligente, hardware genérico y arquitectura distribuida. OneFS no solo es el sistema operativo, sino también el sistema de archivos subyacente que impulsa y almacena los datos en el clúster.

Nodos PowerScale

OneFS funciona exclusivamente con los nodos de plataforma dedicada, lo que se denomina “clúster”. Un solo clúster se compone de varios nodos, que son dispositivos empresariales montables en rack con memoria, CPU, redes, interconexiones Ethernet o InfiniBand de baja latencia, controladoras de disco y medios de almacenamiento. Por lo tanto, cada nodo en el clúster distribuido tiene funcionalidades de computación, almacenamiento y capacidad.

Con la arquitectura Gen6, se requiere un solo chasis de 4 nodos en un factor de forma de 4RU (unidades de rack) para crear un clúster, lo cual escala hasta 252 nodos en OneFS 8.2 y posteriores. Las plataformas de nodos individuales necesitan un mínimo de tres nodos y 3RU de espacio en rack para formar un clúster. Hay distintos tipos de nodos, los cuales se pueden integrar en un solo clúster donde los diferentes nodos proporcionan distintas tasas de capacidad para el rendimiento o las operaciones de entrada/salida por segundo (IOPS). Tanto el chasis Gen6 tradicional como los nodos todo flash F900, F600 y F200 independientes de PowerScale coexistirán felizmente dentro del mismo clúster.

Cada nodo o chasis que se agrega a un clúster aumenta la capacidad del disco, la memoria caché, el CPU y la red. OneFS aprovecha cada uno de los componentes básicos de hardware, por lo que el todo es mayor que la suma de las partes. La RAM se agrupa en una sola caché coherente, lo que permite que las operaciones de I/O de cualquier parte del clúster se beneficien con el almacenamiento de datos en caché en cualquier lugar. Un registro de sistema de archivos garantiza que las operaciones de escritura estén seguras en todas las fallas de alimentación. Los ejes y el CPU se combinan para aumentar el rendimiento, la capacidad y las IOPS a medida que crece el clúster con el fin de obtener acceso a un archivo o a varios archivos. La capacidad de almacenamiento de un clúster puede variar desde decenas de TB a decenas de PB. La capacidad máxima continuará aumentando a medida que los medios de almacenamiento y el chasis de nodos se vuelvan más densos.

Los nodos de la plataforma con tecnología OneFS se dividen en varias clases, o niveles, según su funcionalidad:

Tier	I/O Profile	Drive Media	Nodes	
Performance	High Perf, Low Latency	Flash NVMe/SAS	F900 F600 F200	F810 F800
Hybrid / Utility	Concurrency & Streaming Throughput	SATA/SAS & SSD	H700 H7000	H600 H5600 H500 H400
Archive	Nearline & Deep Archive	SATA	A300 A3000	A200 A2000

Tabla 1: tipos de nodos y niveles de hardware

Red

Hay dos tipos de redes asociadas a un clúster: internas y externas.

Red de back-end

Toda la comunicación dentro de un nodo en un clúster se realiza en una red de back-end exclusiva, que incluye Ethernet o InfiniBand QDR de baja latencia (IB) de 10, 40 o 100 Gb. Esta red de back-end, que está configurada con switches redundantes para la alta disponibilidad, funciona como el backplane para el clúster. Esto permite a cada nodo actuar como un colaborador en el clúster y aislar la comunicación entre nodos a una red privada, de alta velocidad y de baja latencia. Esta red de back-end utiliza el protocolo de Internet (IP) para la comunicación entre nodos.

Red de front-end

Los clientes se conectan al clúster mediante conexiones Ethernet (10 GbE, 25 GbE, 40 GbE o 100 GbE) que están disponibles en todos los nodos. Dado que cada nodo proporciona sus propios puertos Ethernet, la cantidad de ancho de banda de red disponible para el clúster escala linealmente con rendimiento y capacidad. El clúster admite protocolos de comunicación de red estándares para una red del cliente, incluidos NFS, SMB, HTTP, FTP, HDFS y S3. Además, OneFS proporciona una integración completa con entornos IPv4 e IPv6.

Vista completa del clúster

El clúster completo se combina con hardware, software y redes en la siguiente vista:

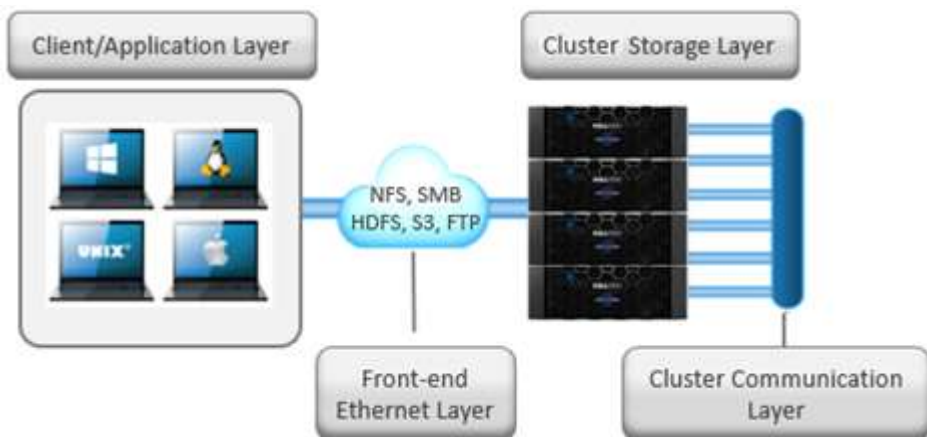


Figura 2: todos los componentes de OneFS en acción

El diagrama anterior muestra la arquitectura completa; el software, el hardware y la red funcionan juntos en el entorno con servidores para proporcionar un solo sistema de archivos completamente distribuido que puede escalar de manera dinámica a medida que cambian las cargas de trabajo y las necesidades de rendimiento y capacidad en un entorno de escalamiento horizontal.

OneFS SmartConnect es un balanceador de carga que funciona en la capa de Ethernet de front-end para distribuir uniformemente las conexiones de clientes en todo el clúster. SmartConnect admite la conmutación por error y la conmutación por recuperación dinámicas de NFS para clientes Linux y UNIX, y la disponibilidad continua de SMB3 para clientes Windows. Esto garantiza que, cuando se produce la falla de un nodo u ocurre un mantenimiento preventivo, todas las lecturas y las escrituras en transferencia se trasladan a otro nodo del clúster para completar su operación sin ninguna interrupción para los usuarios o las aplicaciones.

Durante una falla, los clientes se distribuyen de manera equitativa entre todos los nodos restantes del clúster, garantizando un impacto mínimo en el rendimiento. Si un nodo deja de funcionar por cualquier motivo, incluida una falla, las direcciones IP virtuales de ese nodo se migran sin problemas a otro nodo del clúster. Cuando el nodo fuera de línea vuelve a ponerse en línea, SmartConnect rebalancea automáticamente los clientes NFS y SMB3 en todo el clúster para garantizar el máximo nivel de utilización de almacenamiento y rendimiento. En el caso del mantenimiento periódico del sistema y las actualizaciones de software, esta funcionalidad permite realizar actualizaciones graduales por nodo, lo cual proporciona una disponibilidad completa durante la ventana de mantenimiento.

 Encontrará más información disponible en la documentación técnica de [OneFS SmartConnect](#).

Visión general del software OneFS

Sistema operativo

OneFS se desarrolla sobre una base de sistema operativo (SO) UNIX fundamentado en BSD. Admite la semántica de Linux/UNIX y Windows de forma nativa, incluidos los enlaces físicos, la eliminación al cierre, el cambio de nombre atómico, las ACL y los atributos extendidos. Utiliza BSD como su sistema operativo base, ya que es un sistema operativo maduro y comprobado, y la comunidad de código abierto se puede aprovechar para la innovación. Desde OneFS 8.2 en adelante, la versión del sistema operativo subyacente es FreeBSD 11.

Servicios de cliente

Los protocolos de front-end que los clientes pueden usar para interactuar con OneFS se denominan servicios de cliente. Consulte la sección [Protocolos compatibles](#) para obtener una lista detallada de los protocolos compatibles. A fin de comprender la manera en que OneFS se comunica con los clientes, dividimos el subsistema de I/O en dos mitades: la mitad superior o el “iniciador” y la mitad inferior o el “participante”. Cada nodo del clúster es un participante de una operación de I/O específica. El nodo al que se conecta el cliente es el iniciador y actúa como “capitán” para toda la operación de I/O. La operación de lectura y escritura se detalla en secciones posteriores

Operaciones de clúster

En una arquitectura en clúster, hay trabajos de clúster que son responsables del estado y el mantenimiento del clúster en sí, ya que todos los trabajos son administrados por el motor de trabajos de OneFS. El motor de trabajos se ejecuta en todo el clúster y es responsable de dividir y abordar las tareas de administración y protección de almacenamiento de gran tamaño. Para lograr esto, reduce una tarea a elementos de trabajo más pequeños y, a continuación, asigna o mapea estas partes del trabajo general a varios subprocesos de trabajo en cada nodo. El progreso se rastrea y se informa durante toda la ejecución del trabajo, y se presentan en un estado y un informe detallados tras la finalización.

El motor de trabajos incluye un sistema integral orientado a la comprobación que permite la pausa y la reanudación de los trabajos, además de la detención y el inicio. El marco de trabajo del motor de trabajos también incluye un sistema de administración de impactos adaptable.

Por lo general, el motor de trabajos ejecuta trabajos como tareas en segundo plano en todo el clúster mediante el uso de recursos y capacidad de repuesto o especialmente reservados. Los trabajos mismos se pueden clasificar en tres clases principales:

Trabajos de mantenimiento del sistema de archivos

Estos trabajos ejecutan el mantenimiento del sistema de archivos en segundo plano y, por lo general, requieren acceso a todos los nodos. Estos trabajos se deben ejecutar en configuraciones predeterminadas y, a menudo, en condiciones de clúster degradado. Los ejemplos incluyen la protección del sistema de archivos y las reconstrucciones de unidades.

Trabajos de soporte de funciones

Los trabajos de soporte de funciones realizan el trabajo que facilita algunas funciones de administración de almacenamiento extendida y, por lo general, solo se ejecutan cuando la función ya está configurada. Los ejemplos incluyen la deduplicación y el escaneo antivirus.

Trabajos de acción de usuario

Estos trabajos los ejecuta directamente el administrador de almacenamiento para lograr algún objetivo de administración de datos. Los ejemplos incluyen eliminaciones de árboles en paralelo y el mantenimiento de permisos.

La siguiente tabla proporciona una lista completa de los trabajos del motor de trabajos expuestos, las operaciones que realizan y sus respectivos métodos de acceso al sistema de archivos:

Nombre de trabajo	Descripción del trabajo	Método de acceso
AutoBalance	Balancea el espacio libre en el clúster.	Unidad + LIN
AutoBalanceLin	Balancea el espacio libre en el clúster.	LIN
AVScan	Trabajo de escaneo de virus que ejecutan los servidores de antivirus.	Árbol
ChangelistCreate	Cree una lista de cambios entre dos instantáneas de SyncIQ consecutivas	Changelist
CloudPoolsLin	Archiva los datos en un proveedor de servicio en la nube de acuerdo con una política de pool de archivos.	LIN
CloudPoolsTreewalk	Archiva los datos en un proveedor de servicio en la nube de acuerdo con una política de pool de archivos.	Árbol
Collect	Recupera espacio de disco que no se pudo liberar debido a que un nodo o una unidad no se encontraban disponibles mientras sufrían diversas condiciones de falla.	Unidad + LIN
ComplianceStoreDelete	Trabajo de recolección de elementos no utilizados del modo de cumplimiento de normas de SmartLock.	Árbol
Dedupe	Desduplica bloques idénticos en el sistema de archivos.	Árbol
DedupeAssessment	Ejecuta una evaluación de prueba de los beneficios de la deduplicación.	Árbol
DomainMark	Asocia una ruta y sus contenidos con un dominio.	Árbol
DomainTag	Asocia una ruta y sus contenidos con un dominio.	Árbol
EsrsMftDownload	Trabajo de transferencia administrada de archivos de ESRS para archivos de licencia.	
FilePolicy	Trabajo eficiente de la política de pool de archivos de SmartPools.	Changelist

Nombre de trabajo	Descripción del trabajo	Método de acceso
FlexProtect	Reconstruye y vuelve a proteger el sistema de archivos para que se recupere de un escenario de falla.	Unidad + LIN
FlexProtectLin	Vuelve a proteger el sistema de archivos.	LIN
FSAalyze	Recopila datos de análisis del sistema de archivos que se utilizan junto con InsightIQ.	Changelist
IndexUpdate	Crea y actualiza un índice eficiente del sistema de archivos para los trabajos FilePolicy y FSAalyze.	Changelist
IntegrityScan	Realiza la verificación y la corrección en línea de cualquier incoherencia en el sistema de archivos.	LIN
LinCount	Analiza y cuenta los inodos lógicos (LIN) del sistema de archivos.	LIN
MediaScan	Escanea las unidades en busca de errores en el nivel de medios.	Unidad + LIN
MultiScan	Ejecuta trabajos de Collect y AutoBalance de manera simultánea.	LIN
PermissionRepair	Corrige permisos de archivos y directorios.	Árbol
QuotaScan	Actualiza la contabilidad de cuotas para los dominios creados en una ruta de directorio existente.	Árbol
SetProtectPlus	Aplica la política de archivos predeterminada. Este trabajo se encuentra deshabilitado si SmartPools está activado en el clúster.	LIN
ShadowStoreDelete	Libera el espacio asociado a un almacén oculto.	LIN
ShadowStoreProtect	Protege los almacenes ocultos a los que un LIN hace referencia con una mayor protección solicitada.	LIN
ShadowStoreRepair	Repara los almacenes ocultos.	LIN
SmartPools	Trabajo que ejecuta y transfiere los datos entre los niveles de nodos dentro del mismo clúster. También ejecuta la funcionalidad de CloudPools si este tiene licencia y está configurado.	LIN
SmartPoolsTree	Aplica las políticas de archivos de SmartPools en un subárbol.	Árbol
SnapRevert	Revierte toda una instantánea al cabezal.	LIN
SnapshotDelete	Libera espacio de disco asociado con las instantáneas eliminadas.	LIN
TreeDelete	Elimina una ruta en el sistema de archivos directamente desde el clúster.	Árbol
Undedupe	Elimina la deduplicación de bloques idénticos en el sistema de archivos.	Árbol
Actualización	Actualiza el clúster en una versión posterior de OneFS.	Árbol
WormQueue	Escanea la línea de espera de LIN de SmartLock	LIN

Figura 1: descripción de los trabajos del motor de trabajos de OneFS

A pesar de que los trabajos de mantenimiento del sistema de archivos se ejecutan de forma predeterminada, ya sea según una programación o en respuesta a un evento específico del sistema de archivos, cualquier trabajo del motor de trabajos se puede administrar mediante la configuración de su nivel de prioridad (en relación con otros trabajos) y su política de impacto.

Una política de impacto puede constar de uno o varios intervalos de impacto, que son bloques de tiempo dentro de una semana determinada. Cada intervalo de impacto puede configurarse para utilizar un solo nivel de impacto predefinido que especifica la cantidad de recursos de clúster que se usarán para una operación de clúster específica. Los niveles de impacto disponibles en el motor de trabajos son los siguientes:

- Paused
- Low
- Media
- Alta:

Este grado de granularidad permite la configuración de intervalos y niveles de impacto por trabajo a fin de garantizar una operación de clúster sin problemas. Además, las políticas de impacto resultante dictan cuándo se ejecuta un trabajo y los recursos que este puede consumir.

Además, los trabajos del motor de trabajos se priorizan en una escala de uno a diez, donde un valor menor significa una prioridad más alta. Este es un concepto similar al de la utilidad de programación de UNIX, "nice".

El motor de trabajos permite la ejecución simultánea de hasta tres trabajos. Esta ejecución de trabajos simultánea se rige por los siguientes criterios:

- Prioridad de trabajo
- Conjuntos de exclusión: trabajos que no se pueden ejecutar en conjunto (es decir, FlexProtect y AutoBalance)
- Estado del clúster: la mayoría de los trabajos no se pueden ejecutar cuando el clúster está en un estado degradado.

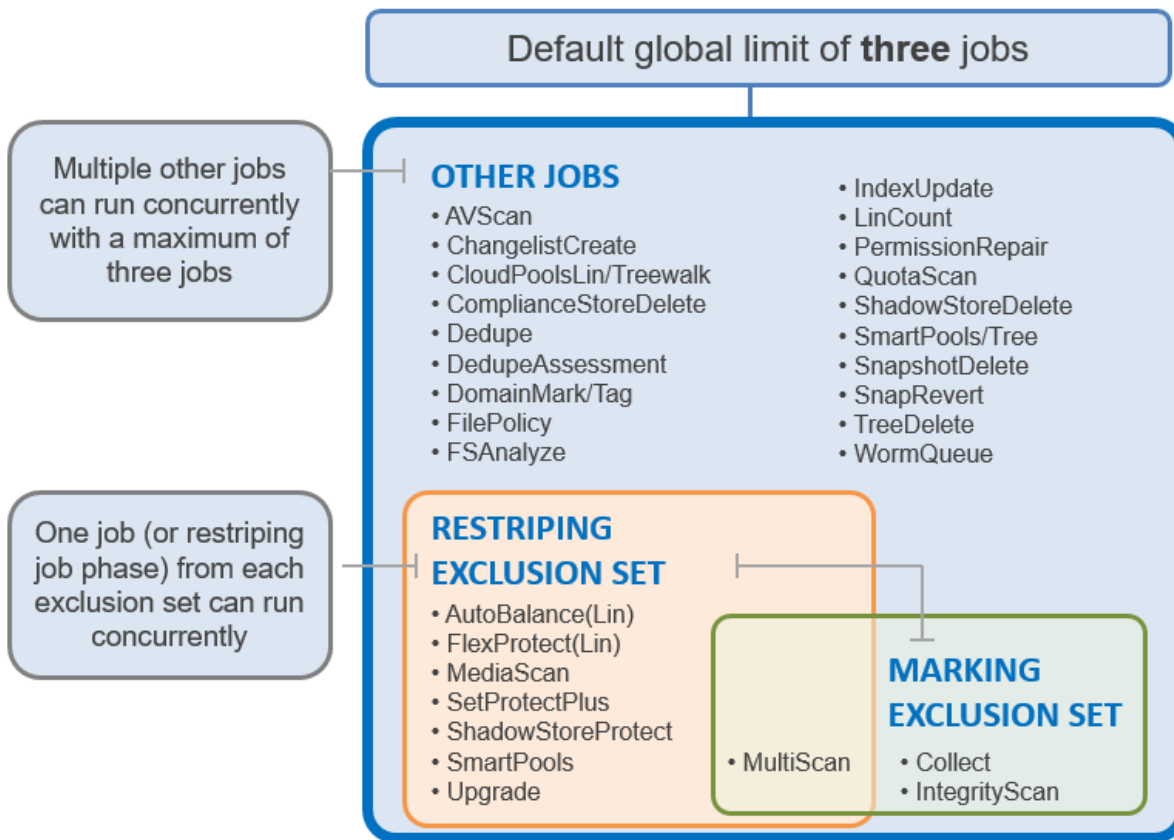


Figura 4: conjuntos de exclusión del motor de trabajos de OneFS

📖 Encontrará más información disponible en la documentación técnica del [motor de trabajos de OneFS](#).

Estructura del sistema de archivos

El sistema de archivos de OneFS se basa en el sistema de archivos de UNIX (UFS) y, por lo tanto, es un sistema de archivos distribuido muy rápido. Cada clúster crea un solo espacio de nombres y un sistema de archivos. Esto significa que el sistema de archivos se distribuye entre todos los nodos del clúster y es accesible para los clientes que se conectan a cualquier nodo del clúster. No hay particionamiento ni es necesario crear un volumen. En lugar de limitar el acceso a espacio libre y a archivos no autorizados en el nivel de volumen físico, OneFS proporciona la misma funcionalidad en software a través de permisos de archivos y recursos compartidos, y a través del servicio SmartQuotas, que proporciona una administración de cuotas en el nivel de directorio.

📖 Encontrará más información disponible en la documentación técnica de [OneFS SmartQuotas](#).

Dado que toda la información se comparte entre los nodos de la red interna, los datos se pueden escribir o leer desde cualquier nodo, lo cual optimiza el rendimiento cuando varios usuarios leen y escriben simultáneamente en el mismo conjunto de datos.

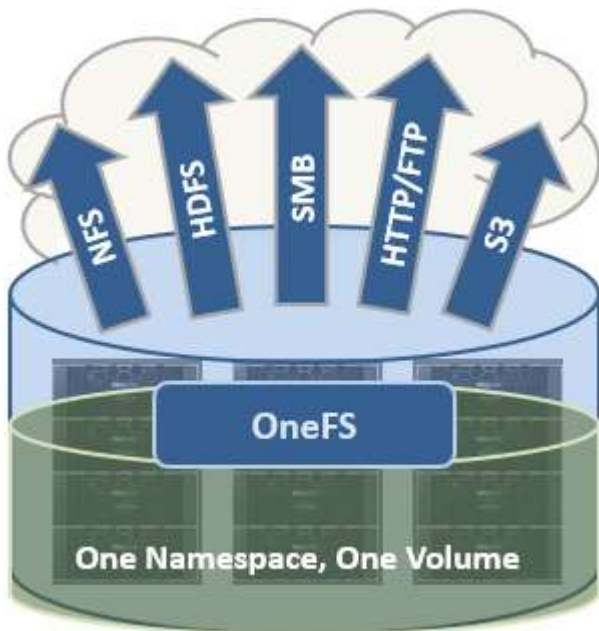


Figura 5: un solo sistema de archivos con varios protocolos de acceso

OneFS es un sistema de archivos único con un espacio de nombres. Los datos y los metadatos se fraccionan en todos los nodos con fines de redundancia y disponibilidad. El almacenamiento se virtualizó completamente para los usuarios y el administrador. El árbol de archivos puede crecer de manera orgánica sin necesidad de planificar ni supervisar la manera en que el árbol crece o la manera en que los usuarios lo utilizan. No es necesario que el administrador aplique un criterio especial para la organización de los archivos en niveles en el disco correspondiente, ya que OneFS SmartPools la manejará automáticamente sin interrumpir el árbol único. Además, no es necesario otorgar ninguna consideración especial a la manera en que se puede replicar un árbol de gran tamaño, ya que el servicio OneFS SyncIQ paraleliza automáticamente la transferencia del árbol de archivos a uno o más clústeres alternativos sin considerar la forma o la profundidad del árbol de archivos.

Este diseño se debe comparar con la agregación de espacios de nombres, que es una tecnología comúnmente utilizada para hacer que el NAS tradicional “aparezca” como un solo espacio de nombres. Con la agregación de espacios de nombres, los archivos todavía deben administrarse en volúmenes independientes, pero una capa simple de “revestimiento” permite que los directorios individuales en los volúmenes se “peguen” a un árbol de “nivel superior” a través de los enlaces simbólicos. En ese modelo, los LUN y los volúmenes, así como los límites de volumen, aún están presentes. Los archivos se deben transferir manualmente de volumen a volumen a fin de balancear la carga. El administrador debe tener cuidado con la manera en que se diseña el árbol. La organización en niveles está lejos de ser transparente y requiere una intervención importante y continua. La conmutación por error requiere el espejeado de archivos entre volúmenes, lo que reduce la eficiencia y aumenta el costo de la compra, la alimentación y el enfriamiento. En general, la carga del administrador al utilizar la agregación de espacios de nombres es más alta que para un dispositivo NAS tradicional simple. Esto evita que esas infraestructuras crezcan a un tamaño muy grande.

Diseño de datos y

OneFS utiliza extensiones y punteros físicos para los metadatos, y almacena los metadatos de archivos y directorios en inodos. Por lo general, los inodos lógicos (LIN) de OneFS tienen un tamaño de 512 bytes, lo que les permite adaptarse a los sectores nativos a los que se formatea la mayoría de los discos duros. También se proporciona compatibilidad con inodos de 8 KB a fin de admitir las clases más densas de disco duro que ahora tienen el formato de sectores de 4 KB.

Los árboles B se utilizan ampliamente en el sistema de archivos, lo que permite la escalabilidad de miles de millones de objetos y búsquedas casi instantáneas de datos o metadatos. OneFS es un sistema de archivos completamente simétrico y altamente distribuido. Los datos y los metadatos siempre son redundantes en varios dispositivos de hardware. Los datos se protegen mediante la codificación de eliminación en los nodos del clúster, lo que crea un clúster de alta eficiencia; esto permite una capacidad de crudo a útil un 80 % mejor en clústeres de cinco nodos o más. Los metadatos (que a menudo constituyen aproximadamente el 1 % del sistema) se espejean en el clúster para aumentar el rendimiento y la disponibilidad. Dado que OneFS no depende de RAID, el administrador puede seleccionar la cantidad de redundancia en el nivel de archivos o directorios más allá de los valores predeterminados del clúster. Las tareas de bloqueo de metadatos y acceso a ellos son administradas por todos los nodos de forma colectiva y equitativa en una arquitectura entre pares. Esta simetría es clave para la sencillez y la resiliencia de la arquitectura. No hay un solo servidor de metadatos, administrador de bloqueos ni nodo de puerta de enlace.

Debido a que OneFS debe acceder a bloques desde varios dispositivos simultáneamente, el esquema de direccionamiento utilizado para los datos y los metadatos se indexa en el nivel físico mediante una tupla de {node, drive, offset}. Por ejemplo, si 12345 era la dirección de un bloque que se alojaba en el disco 2 del nodo 3, entonces se leería {3,2,12345}. Todos los metadatos del clúster se multiplican por espejeado para la protección de datos, al menos en el nivel de redundancia del archivo asociado. Por ejemplo, si un archivo se encuentra en una protección de código de borrado de “+2n”, lo que significa que el archivo podría resistir dos fallas simultáneas, todos los metadatos necesarios para obtener acceso a ese archivo tendrían 3 espejeados, por lo que también podría resistir dos fallas. El sistema de archivos permite inherentemente cualquier estructura para utilizar todos los bloques en cualquier nodo del clúster.

Otros sistemas de almacenamiento envían datos a través de capas de administración de volúmenes y RAID, lo que presenta ineficiencias en el diseño de los datos y proporciona acceso a bloques no optimizados. OneFS controla la ubicación de los archivos directamente, hasta el nivel de sector en cualquier unidad en cualquier lugar del clúster. Esto permite la ubicación optimizada de los datos y los patrones de I/O, y evita las operaciones innecesarias de lectura/modificación/escritura. Mediante la colocación de los datos en los discos archivo por archivo, OneFS es capaz de controlar de manera flexible el tipo de fraccionado, así como el nivel de redundancia del sistema de almacenamiento en los niveles de sistema, directorio e incluso archivo. Los sistemas de almacenamiento tradicionales requerirían que un volumen RAID completo se dedique a un tipo de rendimiento y una configuración de protección específicos. Por ejemplo, un conjunto de discos se puede organizar en una protección RAID 1+0 para una base de datos. Esto dificulta la optimización del uso de ejes en todo el almacenamiento (ya que no es posible tomar prestados los ejes inactivos) y también genera diseños inflexibles que no se adaptan a los requisitos del negocio. OneFS permite ajustes individuales y cambios flexibles en cualquier momento y completamente en línea.

Escritura de archivos

El software OneFS se ejecuta en todos los nodos de manera equitativa, lo que crea un único sistema de archivos que se ejecuta en cada nodo. Ningún nodo controla ni “domina” el clúster; todos los nodos son verdaderos pares.

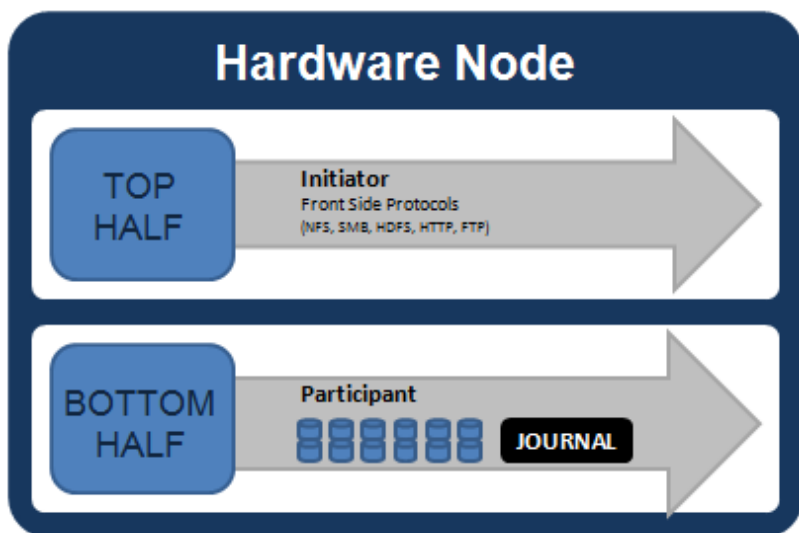


Figura 6: modelo de componentes de nodos involucrados en las actividades de I/O

Si analizáramos todos los componentes dentro de cada nodo de un clúster que participan en las actividades de I/O provenientes de un nivel alto, sería similar a la figura 6. Dividimos la pila en una capa “superior”, denominada iniciador, y una capa “inferior”, denominada participante. Esta división se utiliza como un “modelo lógico” para el análisis de cualquier lectura o escritura determinada. En un nivel físico, los CPU y la caché de RAM en los nodos manejan simultáneamente las tareas de iniciador y participante para las actividades de I/O que ocurren en todo el clúster. Hay cachés y un administrador de bloqueos distribuidos que se excluyen del diagrama anterior para mantener su sencillez. Se analizarán en secciones posteriores del informe.

Cuando un cliente se conecta a un nodo para escribir un archivo, se conecta a la mitad superior o al iniciador de ese nodo. Los archivos se dividen en fragmentos lógicos más pequeños denominados fracciones antes de escribirse en la mitad inferior o en el participante de un nodo (disco). La operación de colocación en el búfer a prueba de fallas mediante un aglomerador de escrituras se utiliza para garantizar la eficiencia de las escrituras y evitar las operaciones de lectura/modificación/escritura. El tamaño de cada fragmento de archivo se conoce como el tamaño de unidad de fracción.

OneFS fracciona los datos en todos los nodos (y no solo en los discos) y protege los archivos, los directorios y los metadatos asociados mediante el código de eliminación de software o la tecnología de espejado. Para los datos, OneFS puede utilizar (según el criterio del administrador) el sistema de codificación de eliminación Reed-Solomon para la protección de datos o el espejado (menos común). El espejado, cuando se aplica a los datos de usuario, tiende a usarse más para casos de rendimiento de alta cantidad de transacciones. Por lo general, la mayor parte de los datos de usuario usará la codificación de eliminación, ya que proporciona un rendimiento extremadamente alto sin sacrificar la eficiencia en disco. La codificación de eliminación puede proporcionar más del 80 % de eficiencia en los discos crudos con cinco nodos o más, y en los clústeres de gran tamaño incluso puede hacerlo mientras proporciona redundancia de un nivel cuatro veces mayor. El ancho de fracción de cualquier archivo dado es la cantidad de nodos (no unidades) en los que se escribe un archivo. Se determina según la cantidad de nodos en el clúster, el tamaño del archivo y la configuración de protección (por ejemplo, +2n).

OneFS utiliza algoritmos avanzados para determinar el diseño de datos a fin de obtener el máximo nivel de eficiencia y rendimiento. Cuando un cliente se conecta a un nodo, el iniciador de ese nodo actúa como el “capitán” para el diseño de datos de escritura de ese archivo. Los datos, la protección del código de eliminación (ECC), los metadatos y los inodos se distribuyen en múltiples nodos dentro de un clúster e incluso en múltiples unidades dentro de los nodos.

La figura 7 a continuación muestra una escritura de archivos que se lleva a cabo en todos los nodos de un clúster de tres nodos.

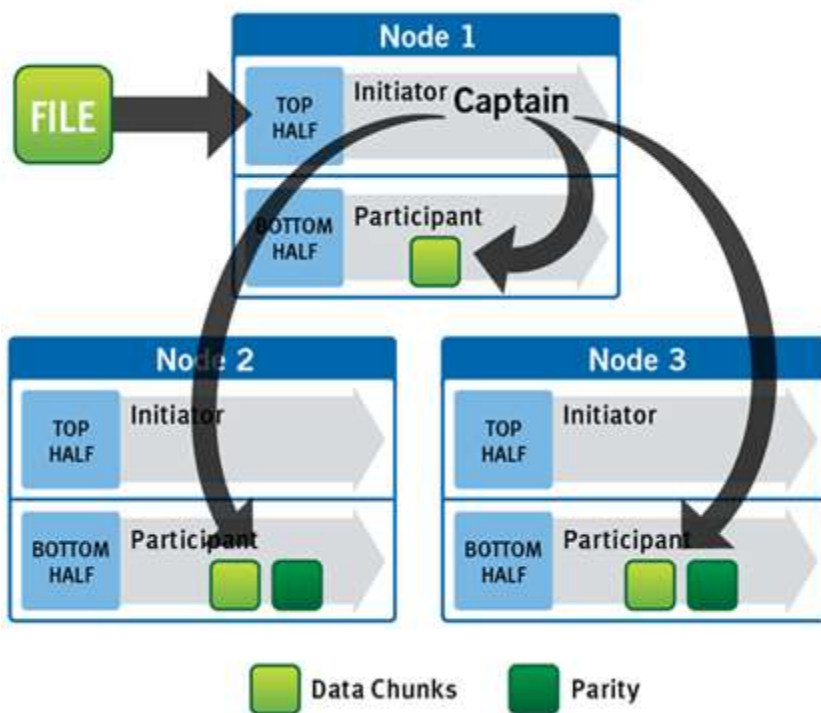


Figura 7: una operación de escritura de archivos en un clúster de 3 nodos

OneFS utiliza la red de back-end para asignar y fraccionar los datos en todos los nodos del clúster de manera automática, por lo que no se requiere ningún procesamiento adicional. A medida que se escriben los datos, se protegen en el nivel especificado. Cuando se realizan escrituras, OneFS divide los datos en unidades atómicas denominadas grupos de protección. La redundancia está integrada en los grupos de protección, de manera que si cada grupo de protección es seguro, todo el archivo está protegido. Para los archivos protegidos por códigos de eliminación, un grupo de protección consta de una serie de bloques de datos, así como un conjunto de códigos de eliminación para esos bloques de datos; para los archivos espejados, un grupo de protección consta de todos los espejados de un conjunto de bloques. OneFS puede cambiar el tipo de grupo de protección que se utiliza en un archivo de manera dinámica, como sería el de escritura. Esto puede permitir muchas funcionalidades adicionales, como que el sistema continúe sin bloqueos en situaciones en las que las fallas temporales de nodos en el clúster impiden el uso de la cantidad deseada de códigos de eliminación. El espejado se puede usar de manera temporal en estos casos para permitir que las escrituras continúen. Cuando se restauran nodos en el clúster, estos grupos de protección espejados se convierten nuevamente de manera transparente y automática a la protección por código de eliminación, sin la intervención del administrador.

El tamaño de bloque del sistema de archivos de OneFS es de 8 KB. Un archivo menor que 8 KB utilizará un bloque completo de 8 KB. Según el nivel de protección de datos, este archivo de 8 KB podría acabar utilizando más de 8 KB de espacio de datos. Sin embargo, la configuración de protección de datos se analiza en detalle en una sección posterior de este informe. OneFS admite sistemas de archivos con miles de millones de archivos pequeños con un rendimiento muy alto, ya que todas las estructuras en disco están diseñadas para escalar a estos tamaños y proporcionar acceso casi instantáneo a cualquier objeto, independientemente de la cantidad total de objetos. Para los archivos de mayor tamaño, OneFS puede aprovechar el uso de varios bloques contiguos de 8 KB. En estos casos, se pueden fraccionar hasta dieciséis bloques contiguos en el disco de un solo nodo. Si un archivo tiene un tamaño de 32 KB, se utilizarán cuatro bloques de 8 KB contiguos.

Para los archivos incluso más grandes, OneFS puede maximizar el rendimiento secuencial aprovechando una unidad de fraccionado que consta de 16 bloques contiguos, para alcanzar un total de 128 KB por unidad de fracción. Durante una escritura, los datos se dividen en unidades de fracción y estos se distribuyen entre varios nodos como un grupo de protección. A medida que los datos se distribuyen en todo el clúster, los códigos de eliminación o los espejados, según sea necesario, se distribuyen dentro de cada grupo de protección para garantizar que los archivos estén protegidos en todo momento.

Una de las funciones clave de la funcionalidad AutoBalance de OneFS es reasignar y rebalancear los datos, y hacer que el espacio de almacenamiento sea más utilizable y eficiente cuando sea posible. En la mayoría de los casos, se puede aumentar el ancho de fracción de archivos más grandes para aprovechar el nuevo espacio libre (a medida que se agregan nodos) y hacer que el fraccionado en disco sea más eficiente. AutoBalance mantiene una alta eficiencia en disco y elimina los “puntos problemáticos” del disco automáticamente.

La mitad superior de iniciador del nodo “capitán” utiliza una transacción de confirmación de dos fases modificada para distribuir de forma segura las escrituras a múltiples NVRAM en todo el clúster, como se muestra en la figura 8 a continuación.

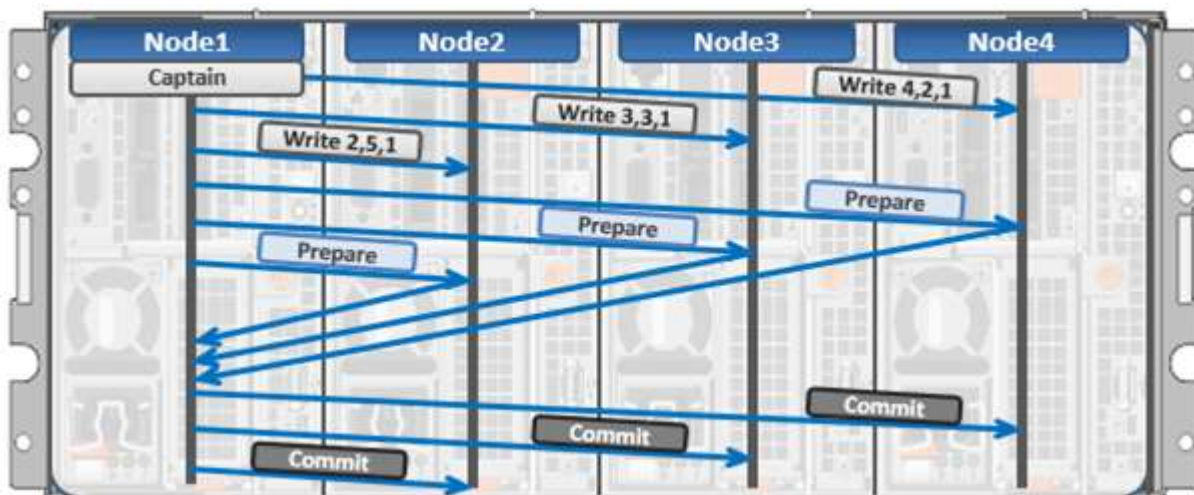


Figura 8: transacciones distribuidas y confirmación de dos fases

Cada nodo que posee bloques en una escritura específica está involucrado en una confirmación de dos fases. El mecanismo depende de NVRAM para el registro de todas las transacciones que ocurren en todos los nodos del clúster de almacenamiento. El uso de múltiples NVRAM en paralelo permite escrituras de alto rendimiento y mantiene los datos protegidos contra todas las fallas, incluidas las fallas de alimentación. En el caso de que un nodo falle en medio de la transacción, esta se reiniciará al instante sin la intervención de ese nodo. Cuando el nodo vuelva, las únicas acciones requeridas serán que el nodo reproduzca su registro desde NVRAM, lo cual tarda segundos o minutos, y que AutoBalance rebalancee ocasionalmente los archivos involucrados en la transacción. No se requieren procesos “fsck” ni “disk-check” costosos. No es necesario realizar ninguna resincronización prolongada. Las escrituras no se bloquean debido a una falla. Este sistema de transacciones patentadas es una de las maneras en las que OneFS elimina los puntos de falla únicos, e incluso varios.

En una operación de escritura, el iniciador “capitanea” o coordina el diseño de los datos y los metadatos, la creación de códigos de eliminación, y las operaciones normales de administración de bloqueo y control de permisos. Un administrador de la interfaz de administración web o CLI en cualquier punto puede optimizar las decisiones de diseño tomadas por OneFS para adaptarse mejor al flujo de trabajo. El administrador puede elegir entre los patrones de acceso a continuación por archivo o en el nivel de directorio:

- **Concurrencia:** optimiza la carga actual en el clúster, con muchos clientes simultáneos. Esta configuración proporciona el mejor comportamiento para las cargas de trabajo mixtas.
- **Streaming:** optimiza el streaming de alta velocidad de un solo archivo, por ejemplo, para permitir una lectura muy rápida con un solo cliente.
- **Aleatorio:** optimiza el acceso impredecible al archivo mediante el ajuste del fraccionado y la deshabilitación del uso de cualquier caché de búsqueda previa.

OneFS también incluye una búsqueda previa adaptable en tiempo real, lo que proporciona el rendimiento de lectura óptimo para los archivos con un patrón de acceso reconocible, sin ninguna intervención administrativa.

❶ El tamaño de archivo más grande compatible actualmente con OneFS aumenta a 16 TB en OneFS 8.2.2 y versiones posteriores, desde un máximo de 4 TB en versiones anteriores.

Almacenamiento en caché de OneFS

El diseño de la infraestructura de almacenamiento en caché de OneFS se basa en la agregación de la caché presente en cada nodo de un clúster en un pool de memoria globalmente accesible. Para ello, OneFS utiliza un sistema de mensajería eficiente, similar al acceso a memoria no uniforme (NUMA). Esto permite que la caché de memoria de todos los nodos esté disponible para cada nodo del clúster. Se accede a la memoria remota a través de una interconexión interna y tiene una latencia mucho menor que la del acceso a las unidades de disco duro.

Para el acceso remoto a la memoria, OneFS utiliza una red Ethernet plana redundante y infrasuscrita como, esencialmente, un bus de sistema distribuido. Aunque no sea tan rápido como la memoria local, el acceso remoto a la memoria sigue siendo muy rápido debido a la baja latencia de Ethernet de 40 Gb.

El subsistema de almacenamiento en caché de OneFS es coherente en todo el clúster. Esto significa que, si existe el mismo contenido en las cachés privadas de múltiples nodos, estos datos almacenados en caché son coherentes en todas las instancias. OneFS utiliza el protocolo MESI para mantener la coherencia de caché. Este protocolo implementa una política de “invalidación durante la escritura” a fin de garantizar que todos los datos sean coherentes en toda la caché compartida.

OneFS utiliza hasta tres niveles de caché de lectura, además de una caché de escritura con respaldo NVRAM, o un aglomerador. Estos, y su interacción de alto nivel, se muestran en el siguiente diagrama.

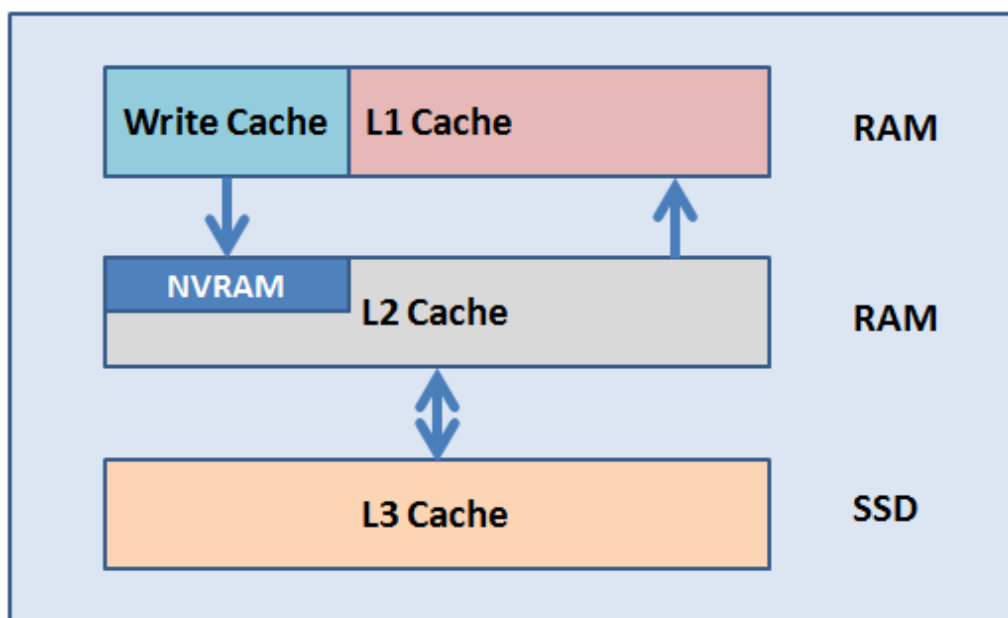


Figura 9: jerarquía de almacenamiento en caché de OneFS

Los primeros dos tipos de caché de lectura, nivel 1 (L1) y nivel 2 (L2), están basados en la memoria (RAM) y son análogos a la caché que se usa en los procesadores (CPU). Estas dos capas de caché están presentes en todos los nodos de almacenamiento de la plataforma.

Nombre	Tipo	Persistencia	Descripción
Caché L1	RAM	Volátil	También llamada caché de front-end, contiene copias limpias y coherentes en el clúster de datos del sistema de archivos y bloques de metadatos solicitados por medio de clientes a través de la red de front-end
Caché L2	RAM	Volátil	Caché de back-end que contiene copias limpias de metadatos y datos del sistema de archivos en un nodo local
SmartCache/ aglomerador de escrituras	NVRAM	No volátil	Caché de registro NVRAM persistente con batería de reserva que coloca en el búfer cualquier escritura pendiente en archivos de front-end que no se han confirmado en el disco.
SmartFlash Caché L3	Disco SSD	No volátil	Contiene bloques de metadatos y datos en archivos liberados de la caché L2, lo cual aumenta eficazmente la capacidad de la caché L2.

Coherencia de caché de OneFS

El subsistema de almacenamiento en caché de OneFS es coherente en todo el clúster. Esto significa que, si existe el mismo contenido en las cachés privadas de múltiples nodos, estos datos almacenados en caché son coherentes en todas las instancias. Por ejemplo, considere el siguiente estado inicial y la secuencia de eventos:

1. El nodo 1 y el nodo 5 tienen una copia de los datos ubicados en una dirección de la caché compartida.
2. El nodo 5, en respuesta a una solicitud de escritura, invalida la copia del nodo 1.
3. Después, el nodo 5 actualiza el valor. (Consulte a continuación).
4. El nodo 1 debe volver a leer los datos de la caché compartida para obtener el valor actualizado.

OneFS utiliza el protocolo MESI para mantener la coherencia de caché. Este protocolo implementa una política de “invalidación durante la escritura” a fin de garantizar que todos los datos sean coherentes en toda la caché compartida. En el siguiente diagrama se ilustran los diversos estados que pueden adoptar los datos en la caché y las transiciones entre ellos. Los distintos estados de la figura son los siguientes:

- M: modificados: los datos solo existen en la caché local y su valor cambió a partir del valor en la caché compartida. Por lo general, los datos modificados se conocen como defectuosos.
- E: exclusivos: los datos solo existen en la caché local, pero coinciden con lo que se encuentra en la caché compartida. A menudo, estos datos se conocen como limpios.
- S: compartidos: los datos de la memoria caché local también pueden estar en otras cachés locales del clúster.
- I: no válidos: se perdió un bloqueo (exclusivo o compartido) en los datos.

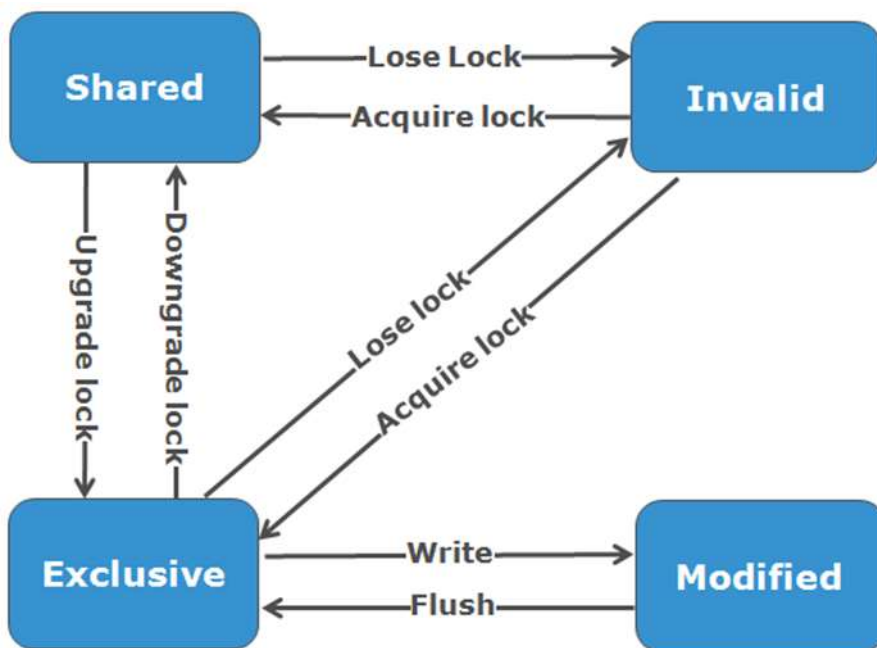


Figura 10: diagrama del estado de la coherencia de caché de OneFS

Caché de nivel 1

La caché de nivel 1 (L1) o de front-end es la memoria más cercana a las capas de protocolo (por ejemplo, NFS, SMB, etc.) que utilizan los clientes o los iniciadores conectados a ese nodo. El propósito principal de la caché L1 es buscar previamente los datos en los nodos remotos. Los datos se buscan previamente por archivo, y esto está optimizado para reducir la latencia asociada con la red de back-end de los nodos. Dado que la latencia de interconexión de back-end es relativamente pequeña, el tamaño de la caché L1 y la cantidad típica de datos almacenados por solicitud son inferiores a la caché L2.

L1 también se conoce como caché remota, ya que contiene los datos recuperados de otros nodos del clúster. Es coherente en todo el clúster, pero solo la usa el nodo en el que reside y no es accesible para otros nodos. Los datos de la caché L1 en los nodos de almacenamiento se descartan enérgicamente después de su uso. La caché L1 utiliza el direccionamiento basado en archivos, en la que se accede a los datos a través de un desplazamiento en un objeto de archivo.

La caché L1 se refiere a la memoria en el mismo nodo que el iniciador. Solo el nodo local puede acceder a ella y, por lo general, la caché no es la copia maestra de los datos. Esto es similar a la caché L1 en un núcleo de CPU, que se puede invalidar a medida que otros núcleos escriben en la memoria principal.

La coherencia de caché L1 se administra a través de un protocolo similar a MESI mediante bloqueos distribuidos, como se describe anteriormente.

OneFS también utiliza una caché de inodo exclusiva en la cual se conservan los inodos solicitados recientemente. Con frecuencia, la caché de inodo tiene un gran impacto en el rendimiento, ya que los clientes suelen almacenar datos en la caché, y muchas actividades de I/O de red son principalmente solicitudes de atributos de archivos y metadatos que se pueden devolver rápidamente desde el inodo almacenado en caché.

① La caché L1 se utiliza de manera diferente en los nodos aceleradores del clúster, que no contienen ninguna unidad de disco. En lugar de eso, toda la caché de lectura es L1, ya que todos los datos se recuperan desde otros nodos de almacenamiento. Además, la antigüedad de la caché se basa en una política de expulsión de recursos menos usados recientemente (LRU), en lugar del algoritmo de descarga que generalmente se utiliza en la caché L1 de un nodo de almacenamiento. Debido a que la caché L1 de un acelerador es grande, los datos en ella tienen muchas más posibilidades de solicitarse nuevamente, por lo que los bloques de datos no se eliminan inmediatamente de la caché cuando se usan. Sin embargo, los metadatos y las cargas de trabajo con actualizaciones intensivas no se benefician mucho, y la caché de un acelerador solo es beneficiosa para los clientes conectados directamente al nodo.

Caché de nivel 2

La caché de nivel 2 (L2), o caché de back-end, se refiere a la memoria local del nodo en el que se almacena un bloque de datos en especial. La caché L2 es accesible globalmente desde cualquier nodo del clúster y se utiliza para reducir la latencia de una operación de lectura, ya que no se requiere una búsqueda directa de las unidades de disco. Por lo tanto, la cantidad de datos que se buscan previamente en la caché L2 para su uso por parte de nodos remotos es mucho mayor que la de la caché L1.

La caché L2 también se conoce como caché local, ya que contiene los datos recuperados de las unidades de disco que se encuentran en ese nodo, los que luego se ponen a disposición de las solicitudes de los nodos remotos. Los datos en la caché L2 se expulsan de acuerdo con un algoritmo de menos usados recientemente (LRU).

El nodo local aborda los datos en la caché L2 mediante una compensación en una unidad de disco que es local para ese nodo. Dado que el nodo sabe dónde se encuentran los datos solicitados por los nodos remotos en el disco, esta es una manera muy rápida de recuperar los datos destinados a los nodos remotos. Un nodo remoto accede a la caché L2 mediante una búsqueda de la dirección de bloque para un objeto de archivo específico. Como se describió anteriormente, no hay ninguna invalidación de MESI necesaria aquí y la caché se actualiza automáticamente durante las escrituras y se mantiene coherente con el sistema de transacciones y la NVRAM.

Caché de nivel 3

También se puede configurar un tercer nivel de caché de lectura opcional, denominado SmartFlash o caché de nivel 3 (L3), en los nodos que contienen unidades de estado sólido (SSD). SmartFlash (L3) es una caché de expulsión que se completa con bloques de caché L2 a medida que se vuelven obsoletos en la memoria. Utilizar unidades SSD para el almacenamiento en caché, en vez usarlos como dispositivos tradicionales de almacenamiento de sistemas de archivos, conlleva beneficios significativos. Por ejemplo, cuando se reserva para el almacenamiento en caché, se utiliza toda la unidad SSD y las operaciones de escritura se realizan de manera muy lineal y predecible. Esto da como resultado una mejor utilización y también un desgaste considerablemente menor, así como una mayor durabilidad en comparación con el uso del sistema de archivos periódico, en particular con cargas de trabajo de escritura aleatoria. El uso de SSD para caché también hace que el dimensionamiento de la capacidad de SSD sea un panorama mucho más sencillo y menos propenso a errores en comparación con el uso de SSD como nivel de almacenamiento.

En el siguiente diagrama se ilustra la manera en que los clientes interactúan con la infraestructura de la caché de lectura de OneFS y el aglomerador de escrituras. La caché L1 continúa interactuando con la caché L2 en cualquier nodo que lo requiera, y la caché L2 interactúa con el subsistema de almacenamiento y la caché L3. La caché L3 se almacena en una unidad SSD dentro del nodo, y cada nodo en el mismo pool de nodos tiene habilitada la caché L3.

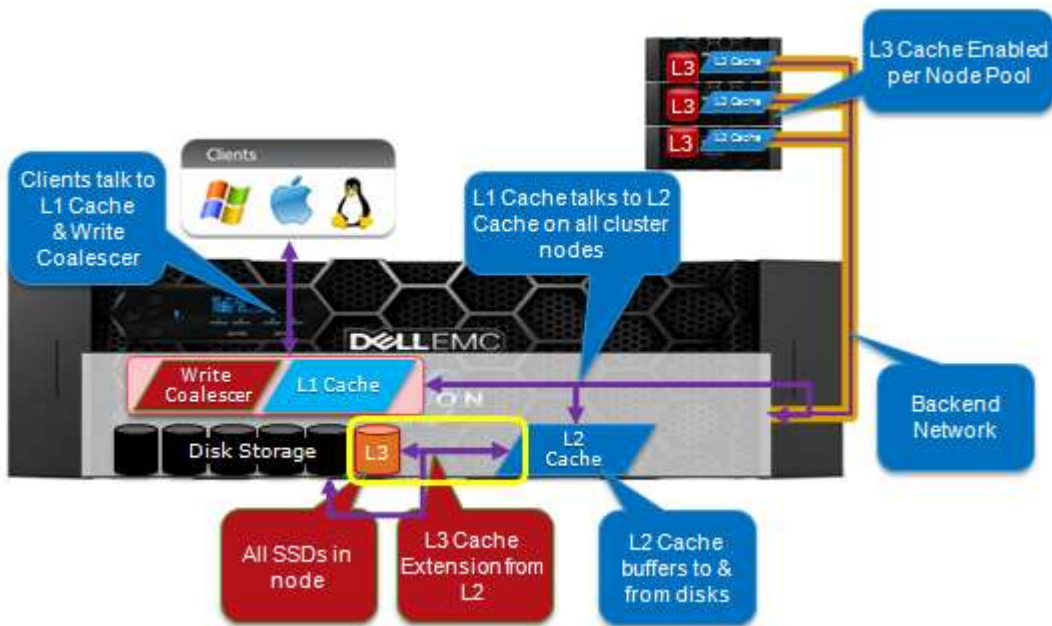


Figura 11: arquitectura de almacenamiento en caché L1, L2 y L3 de OneFS

OneFS indica que un archivo se escribe en varios nodos del clúster y posiblemente en varias unidades de un nodo, por lo que todas las solicitudes de lectura implican la lectura de datos remotos (y posiblemente locales). Cuando una solicitud de lectura llega a un cliente, OneFS determina si los datos solicitados están en la caché local. Todos los datos que residen en la caché local se leen de inmediato. Si los datos solicitados no están en la caché local, se leen desde el disco. Para los datos que no están en el nodo local, se realiza una solicitud desde los nodos remotos en los cuales residen. En cada uno de los nodos restantes, se realiza otra búsqueda en la caché. Todos los datos en la caché se devuelven de inmediato, y todos los datos que no se encuentren en la caché se recuperan desde el disco.

Cuando los datos se recuperan desde la caché local y remota (y posiblemente del disco), regresan al cliente.

Los pasos de alto nivel para completar una solicitud de lectura en un nodo local y remoto son los siguientes:

En el nodo local (el nodo que recibe la solicitud):

1. Determine si parte de los datos solicitados se encuentran en la caché L1 local. De ser así, regrese al cliente.
2. Si no se encuentran en la caché local, solicite datos desde los nodos remotos.

En los nodos remotos:

1. Determine si los datos solicitados se encuentran en la caché L2 o L3 local. Si es así, regrese al nodo solicitante.
2. Si no se encuentra en la caché local, lea desde el disco y regrese al nodo solicitante.

La caché de escritura acelera el proceso de escribir datos en un clúster. Esto se logra mediante la organización de solicitudes de escritura más pequeñas en lotes y su envío al disco en fragmentos más grandes, lo que elimina una cantidad significativa de latencia de escritura de disco. Cuando los clientes escriben en el clúster, OneFS escribe temporalmente los datos en una caché de registro basada en NVRAM en el nodo del iniciador, en lugar de escribirlos inmediatamente en el disco. A continuación, OneFS puede vaciar estas escrituras almacenadas en caché en el disco en otro momento o cuando sea más conveniente. Además, estas escrituras también se espejean en los registros de NVRAM de los nodos participantes para satisfacer los requisitos de protección del archivo. Por lo tanto, en el caso de una división del clúster o de una interrupción inesperada de los nodos, las escrituras en caché sin confirmar están completamente protegidas.

La caché de escritura funciona de la siguiente manera:

- Un cliente de NFS envía una solicitud de escritura al nodo 1 para un archivo con la protección +2n.
- El nodo 1 acepta las escrituras en la caché de escritura de NVRAM (ruta rápida) y luego espejea las escrituras a los archivos de registro de los nodos participantes para la protección.
- Las confirmaciones de escritura se devuelven al cliente NFS de inmediato, por lo que se evita la latencia de escritura en disco.
- A medida que se llena la caché de escritura del nodo 1, se vacía periódicamente, y las escrituras se confirman en el disco mediante el proceso de asignación de dos fases (descrito anteriormente) con la protección de código de eliminación (ECC) adecuada (+2n).
- Los archivos de registro de la caché de escritura y del nodo de participante se borran y están disponibles para aceptar nuevas escrituras.

 Encontrará más información disponible en la documentación técnica de [OneFS SmartFlash](#).

Lecturas de archivos

Los datos, los metadatos y los inodos se distribuyen en múltiples nodos dentro de un clúster e incluso en múltiples unidades dentro de los nodos. Al leer o escribir en el clúster, el nodo al que se conecta un cliente actúa como el “capitán” de la operación.

En una operación de lectura, el nodo “capitán” recopila todos los datos de los distintos nodos del clúster y los presenta de manera coherente para el solicitante.

Debido al uso de hardware estándar del sector optimizado para el costo, el clúster proporciona una alta relación de caché a disco (múltiples GB por nodo) que se asigna dinámicamente para las operaciones de lectura y escritura según sea necesario. Esta caché basada en RAM está unificada y es coherente en todos los nodos del clúster, lo que permite que la solicitud de lectura de un cliente en un nodo pueda beneficiarse de las operaciones de I/O ya tramitadas en otro nodo. Se puede acceder rápidamente a estos bloques almacenados en caché desde cualquier nodo en el backplane de baja latencia. Esto permite una caché de RAM grande y eficiente, lo que acelera en gran medida el rendimiento de lectura.

A medida que el clúster se hace más grande, aumenta el beneficio de la caché. Por esta razón, la cantidad de I/O en el disco de un clúster suele ser considerablemente menor que la de las plataformas tradicionales, lo que permite reducir las latencias y mejorar la experiencia de usuario.

Para los archivos marcados con un patrón de acceso concurrent o streaming, OneFS puede aprovechar la búsqueda previa de datos en función de la heurística que utiliza el componente SmartRead. SmartRead puede crear una “canalización” de datos desde la caché L2, que se busca previamente en una caché local “L1” en el nodo “capitán”. Esto mejora en gran medida el rendimiento de lectura secuencial en todos los protocolos, lo que significa que las lecturas provienen directamente de la RAM en cuestión de milisegundos. Para los casos altamente secuenciales, SmartRead puede realizar una búsqueda previa con mucho dinamismo, lo que permite lecturas o escrituras de archivos individuales con tasas de datos muy altas.

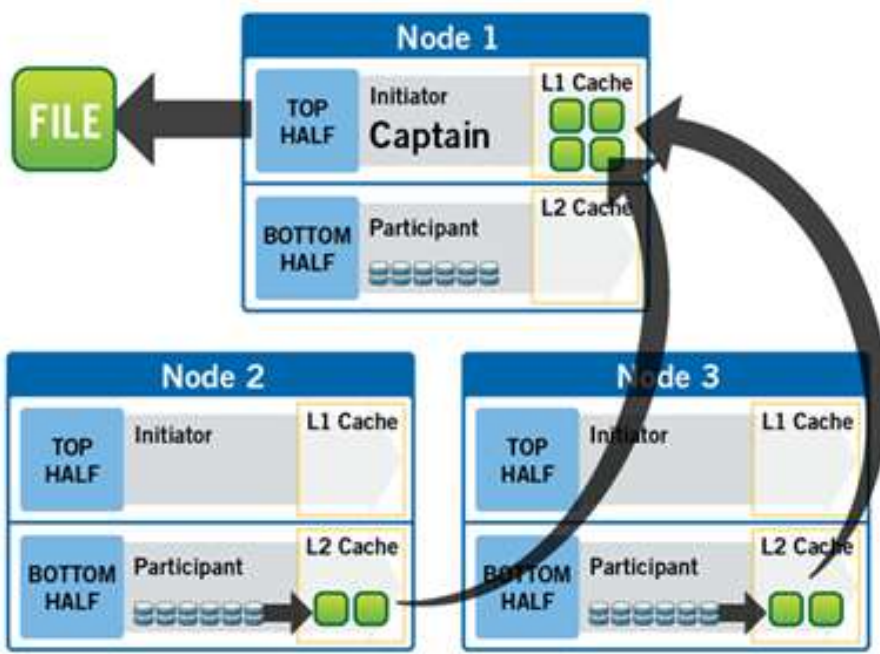


Figura 12: una operación de lectura de archivos en un clúster de 3 nodos

En la figura 10 se ilustra la manera en que SmartRead lee un archivo no almacenado en caché de acceso secuencial que solicita un cliente conectado al nodo 1 en un clúster de 3 nodos.

1. El nodo 1 lee los metadatos para identificar la ubicación donde existen todos los bloques de datos en archivos.
2. El nodo 1 también comprueba su caché L1 para ver si tiene los datos en archivos que se solicitan.
3. El nodo 1 crea una canalización de lectura y envía solicitudes simultáneas a todos los nodos que tienen una pieza de datos en archivos para recuperar esos datos desde el disco.
4. Cada nodo extrae los bloques de datos en archivos del disco a su caché L2 (o caché L3 de SmartFlash cuando está disponible), y transmite los datos en archivos al nodo 1.
5. El nodo 1 registra los datos entrantes a la caché L1 y suministra simultáneamente el archivo al cliente. Mientras tanto, el proceso de búsqueda previa continúa.
6. Para casos altamente secuenciales, los datos en la caché L1 pueden “descartarse” de manera opcional a fin de liberar la RAM para otras exigencias de caché L1 o L2.

El almacenamiento en caché inteligente de SmartRead permite un rendimiento de lectura muy elevado con altos niveles de acceso simultáneo. Cabe destacar que es más rápido que el nodo 1 obtenga datos en archivos de la caché del nodo 2 (mediante la interconexión de clústeres de baja latencia), en vez de obtener acceso a su propio disco local. Los algoritmos de SmartRead controlan la agresividad de la búsqueda previa (deshabilita la búsqueda previa en los casos de acceso aleatorio) y el periodo durante el cual se conservan los datos en la caché, y optimiza la ubicación donde se almacenan los datos en caché.

Bloqueos y simultaneidad

OneFS incluye un administrador de bloqueos distribuidos que coordina los bloqueos de datos en todos los nodos de un clúster de almacenamiento. El administrador de bloqueo es altamente ampliable y permite que múltiples “personalidades” de bloqueo admitan bloqueos del sistema de archivos y bloqueos en el nivel de protocolo coherentes con el clúster, como los bloqueos en modo de recurso compartido SMB o los bloqueos en modo de recomendación de NFS. OneFS también es compatible con bloqueos delegados, como bloqueos oportunos de CIFS y delegaciones de NFSv4.

Cada nodo de un clúster es un coordinador para bloquear recursos, y un coordinador se asigna a los recursos con bloqueos basado en un algoritmo de hash avanzado. La manera en que se diseña el algoritmo es que el coordinador casi siempre termina en un nodo diferente al iniciador de la solicitud. Cuando se solicita un bloqueo para un archivo, puede ser un bloqueo compartido (lo que permite que varios usuarios compartan el bloqueo de forma simultánea, generalmente para lecturas) o un bloqueo exclusivo (lo que permite un usuario en cualquier momento dado, a menudo para escrituras).

La figura 13 a continuación muestra un ejemplo de cómo los subprocesos de distintos nodos pueden solicitar un bloqueo del coordinador.

1. El nodo 2 está designado como el coordinador de estos recursos.
2. El subproceso 1 del nodo 4 y el subproceso 2 del nodo 3 solicitan un bloqueo compartido en un archivo del nodo 2 al mismo tiempo.
3. El nodo 2 comprueba si existe un bloqueo exclusivo para el archivo solicitado.
4. Si no existen bloqueos exclusivos, el nodo 2 otorga al subproceso 1 del nodo 4 y al subproceso 2 del nodo 3 bloqueos compartidos en el archivo solicitado.
5. El nodo 3 y el nodo 4 ahora ejecutan una lectura en el archivo solicitado.
6. El subproceso 3 del nodo 1 solicita un bloqueo exclusivo para el mismo archivo como si el nodo 3 y el nodo 4 lo estuvieran leyendo.
7. El nodo 2 comprueba el nodo 3 y el nodo 4 para ver si se pueden recuperar los bloqueos compartidos.
8. El nodo 3 y el nodo 4 continúan leyendo, por lo que el nodo 2 le pide al subproceso 3 del nodo 1 que espere un breve instante.
9. El subproceso 3 del nodo 1 se bloquea hasta que el nodo 2 otorga el bloqueo exclusivo y, a continuación, finaliza la operación de escritura.

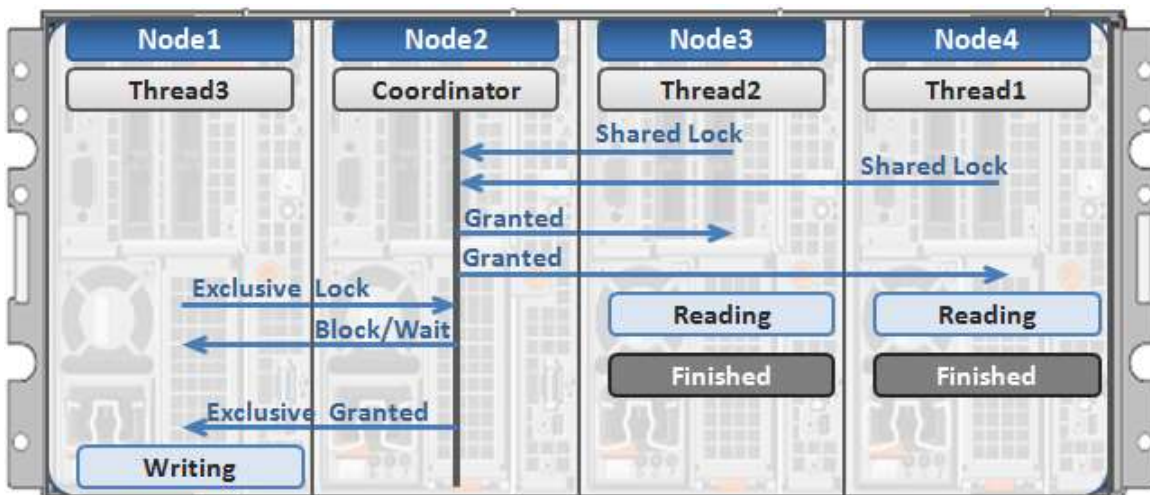


Figura 13: administrador de bloqueo distribuido

I/O multiproceso

Con el creciente uso de almacenes de datos NFS de gran tamaño para la virtualización de servidores y el soporte de aplicaciones empresariales, viene la necesidad de alto rendimiento y baja latencia para los archivos grandes. Para permitir esto, OneFS Multi-writer admite la escritura simultánea de varios subprocesos en archivos individuales.

En el ejemplo anterior, el acceso de escritura simultáneo a un archivo grande puede verse limitado por el mecanismo de bloqueo exclusivo, que se aplica a todo el nivel de archivo. A fin de evitar este posible cuello de botella, OneFS Multi-writer proporciona un bloqueo de escritura más granular mediante la subdivisión del archivo en regiones separadas y la concesión de bloqueos de escritura exclusivos a regiones individuales, en lugar de al archivo completo. Por lo tanto, varios clientes pueden escribir simultáneamente en distintas partes del mismo archivo.

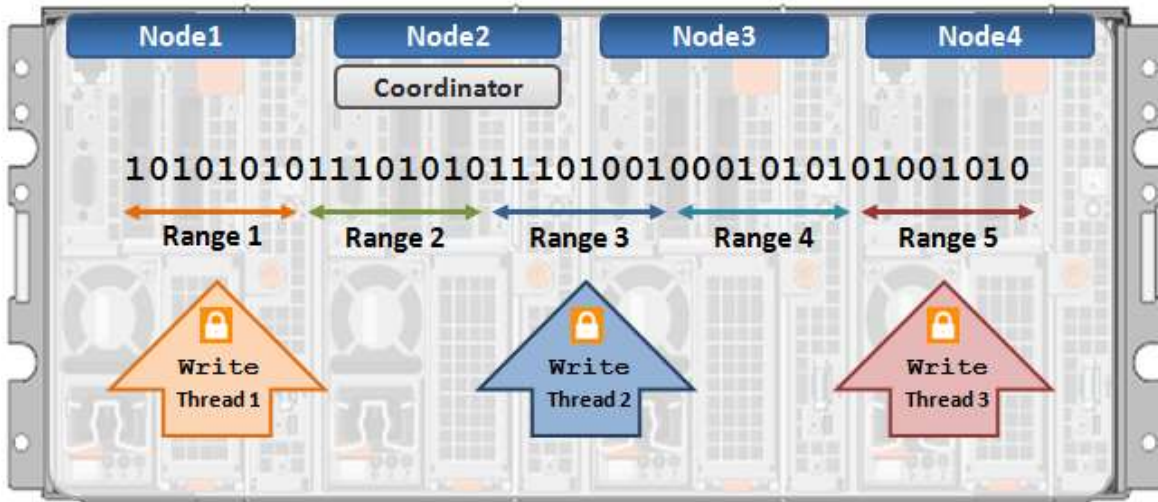


Figura 14: escritor de I/O multiproceso

Protección de datos

Pérdida de alimentación

Un registro de sistema de archivos, que almacena información acerca de los cambios en el sistema de archivos, está diseñado para permitir recuperaciones rápidas y coherentes después de fallas del sistema o de fallas generales, como la pérdida de alimentación. El sistema de archivos reproduce las entradas del registro después de que un nodo o un clúster se recupera de una pérdida de alimentación u otra interrupción. Sin un registro, un sistema de archivos necesitaría examinar y revisar cada cambio potencial individualmente después de una falla (una operación de "fsck" o "chkdsk"); en un sistema de archivos de gran tamaño, esta operación puede tardar mucho tiempo.

OneFS es un sistema de archivos de registro en el cual cada nodo contiene una tarjeta NVRAM con batería de reserva utilizada para proteger las escrituras sin confirmar en el sistema de archivos. La batería de la tarjeta NVRAM dura muchos días sin la necesidad de una recarga. Cuando se inicia un nodo, comprueba su registro y reproduce selectivamente las transacciones al disco si el sistema de registro lo considera necesario.

OneFS se montará solo si puede garantizar el registro de todas las transacciones que aún no se encuentran en el sistema. Por ejemplo, si no se seguían los procedimientos de apagado correctos y se descargaba la batería de NVRAM, se podían perder las transacciones; para evitar posibles problemas, el nodo no montará el sistema de archivos.

Fallas de hardware y quórum

Para que el clúster funcione correctamente y acepte escrituras de datos, debe haber un quórum de nodos activos y con capacidad de respuesta. Un quórum se define como una mayoría sencilla: un clúster con nodos debe tener $\lfloor n/2 \rfloor + 1$ nodos en línea a fin de permitir las escrituras. Por ejemplo, en un clúster de siete nodos, se requerirían cuatro nodos para un quórum. Si un nodo o grupo de nodos está activo y responde, pero no es miembro de un quórum, se ejecuta en un estado de solo lectura.

OneFS utiliza un quórum para evitar las condiciones de “división de recursos” que se pueden introducir si el clúster se debe dividir temporalmente en dos clústeres. Mediante el seguimiento de la regla del quórum, la arquitectura garantiza que, independientemente de la cantidad de nodos que falle o vuelva a estar en línea, si se lleva a cabo una escritura, esta puede ser coherente con cualquier escritura anterior que haya ocurrido. El quórum también define la cantidad de nodos necesaria para cambiarse un nivel de protección de datos determinado. Para un nivel de protección basado en el código de eliminación de +, el clúster debe contener al menos 2+1 nodos. Por ejemplo, se requiere un mínimo de siete nodos para una configuración +3n; esto permite una pérdida simultánea de tres nodos y, al mismo tiempo, mantiene un quórum de cuatro nodos para que el clúster permanezca completamente operativo. Si un clúster se encuentra por debajo del quórum, el sistema de archivos se colocará automáticamente en un estado protegido y de solo lectura, lo que rechazará las escrituras, pero seguirá permitiendo el acceso de lectura a los datos disponibles.

Fallas de hardware: agregar o quitar nodos

Un sistema llamado el protocolo de administración de grupos (GMP) permite un conocimiento global del estado del clúster en todo momento y garantiza una vista coherente en todo el clúster del estado de todos los demás nodos. Si uno o más nodos se vuelven inaccesibles a través de la interconexión del clúster, el grupo se “divide” o se quita del clúster. Todos los nodos se resuelven con una nueva vista coherente de su clúster. (Piense en esto como si el clúster estuviera dividido en dos grupos independientes de nodos, pero tenga en cuenta que solo un grupo puede tener quórum). Mientras se encuentra en este estado dividido, todos los datos del sistema de archivos son accesibles y, del lado que mantiene el quórum, se pueden modificar. Todos los datos almacenados en el dispositivo “inactivo” se reconstruyen mediante la redundancia almacenada en el clúster.

Si el nodo vuelve a estar accesible, se produce una “fusión” o una adición, lo que hace que los nodos vuelvan al clúster. (Los dos grupos vuelven a fusionarse en uno). El nodo puede volver a unirse al clúster sin reconstruirse ni volver a configurarse. Esto difiere de los arreglos RAID de hardware, que requieren la reconstrucción de unidades. AutoBalance puede volver a fraccionar algunos archivos para aumentar la eficiencia, si algunos de sus grupos de protección se sobrescribieron y se transformaron en fracciones más reducidas durante la división.

El motor de trabajos de OneFS también incluye un proceso denominado Collect, que actúa como un recopilador de huérfanos. Cuando un clúster se divide durante una operación de escritura, es posible que algunos bloques asignados para el archivo deban volver a asignarse del lado del quórum. Se trata de bloques asignados “huérfanos” del lado que no corresponde al quórum. Cuando el clúster se vuelve a fusionar, el trabajo de Collect ubicará estos bloques huérfanos a través de un escaneo de marca y barrido paralelo y los recuperará como espacio libre para el clúster.

Reconstrucción escalable

OneFS no se basa en RAID de hardware, ya sea para la asignación de datos o para la reconstrucción de datos después de fallas. En lugar de eso, OneFS administra la protección de los datos en archivos directamente, y cuando se produce una falla, reconstruye los datos de manera paralela. OneFS es capaz de determinar los archivos que se ven afectados por una falla en constante en el tiempo mediante la lectura de los datos del inodo de forma lineal, directamente del disco. El conjunto de archivos afectados se asigna a un conjunto de subprocesos de trabajo que se distribuyen entre los nodos del clúster mediante el motor de trabajos. Los nodos trabajadores reparan los archivos en paralelo. Esto significa que, a medida que el clúster aumenta de tamaño, se reduce el tiempo de reconstrucción debido a fallas. Esto tiene una gran ventaja de eficiencia en el mantenimiento de la resiliencia de los clústeres a medida que aumenta su tamaño.

Hot spare virtual

La mayoría de los sistemas de almacenamiento tradicionales basados en RAID requieren el aprovisionamiento de una o más unidades de “hot spare” para permitir la recuperación independiente de unidades fallidas. La unidad hot spare reemplaza a la unidad fallida en un conjunto RAID. Si estos hot spares no se reemplazan por sí mismos antes de que aparezcan más fallas, el sistema arriesga una pérdida de datos catastrófica. OneFS evita el uso de unidades hot spare y simplemente toma prestado el espacio libre disponible en el sistema a fin de recuperarse de las fallas. Esta técnica se denomina hot spare virtual. Al hacerlo, permite que el clúster se repare automáticamente por completo, sin intervención humana. El administrador puede crear una reserva de hot spare virtual, lo que permite que el sistema se repare a pesar de que los usuarios realizan escrituras continuas.

Protección de datos en el nivel de archivos con codificación de eliminación

Un clúster está diseñado para tolerar una o más fallas de componentes simultáneas sin impedir que el clúster suministre datos. Para lograr esto, OneFS protege los archivos con la protección basada en código de borrado, la corrección de errores Reed-Solomon (protección N+M) o un sistema de espejado. La protección de datos se aplica en forma de software en el nivel de archivos, lo que permite que el sistema se centre en la recuperación de solamente los archivos que se ven afectados por una falla, en lugar de tener que comprobar y reparar un volumen o un conjunto de archivos completo. Los metadatos y los inodos de OneFS cuentan con la protección continua del espejado, en lugar de la codificación Reed-Solomon, y con al menos el mismo nivel de protección que los datos a los que hacen referencia.

Debido a que todos los datos, los metadatos y la información de protección se distribuyen entre los nodos del clúster, el clúster no requiere una unidad o un nodo de paridad exclusivo o un dispositivo o un conjunto de dispositivos dedicados para administrar los metadatos. Esto garantiza que ningún nodo se convierta en un punto único de falla. Todos los nodos comparten equitativamente las tareas que se deben llevar a cabo, lo que proporciona una simetría perfecta y balanceo de carga en una arquitectura entre pares.

OneFS ofrece varios niveles de ajustes de protección de datos configurables, los cuales se pueden modificar en cualquier momento sin necesidad de poner el clúster o el sistema de archivos offline.

Para un archivo protegido con códigos de eliminación, nos referimos a que cada uno de sus grupos de protección está protegido a un nivel de $N+M/b$, donde $N > M$ y $M \geq b$. Los valores N y M representan, respectivamente, la cantidad de unidades que se utilizan para los datos y los códigos de eliminación dentro del grupo de protección. El valor de b se relaciona con la cantidad de fracciones de datos que se usan para distribuir ese grupo de protección, y lo veremos a continuación. Un caso común y fácil de comprender es donde $b=1$, lo que significa que un grupo de protección incorpora: datos equivalentes a N unidades; redundancia equivalente a M unidades, almacenada en códigos de eliminación; y que el grupo de protección se debe disponer en una sola fracción dentro de un conjunto de nodos. Esto permite que M miembros del grupo de protección fallen de forma simultánea y que aún proporcionen una disponibilidad de datos del 100 %. Los M miembros del código de eliminación se calculan a partir de los N miembros de datos. La figura 13 a continuación muestra el caso de un grupo de protección 4+2 normal ($N = 4$, $M = 2$, $b = 1$).

Dado que OneFS fracciona los archivos en todos los nodos, esto implica que los archivos fraccionados en $N+M$ pueden resistir fallas simultáneas de nodos sin pérdida de disponibilidad. Por lo tanto, OneFS proporciona resiliencia ante cualquier tipo de falla, ya sea en una unidad, un nodo o un componente dentro de un nodo (por ejemplo, una tarjeta). Además, un nodo cuenta como una sola falla, independientemente de la cantidad o el tipo de componentes que fallen dentro de él. Por lo tanto, si fallan cinco unidades en un nodo, solo se cuentan como una falla única para los fines de protección $N+M$.

OneFS puede proporcionar de manera exclusiva un nivel variable de M , hasta cuatro, lo que proporciona una protección de falla cuádruple. Esto supera con creces el nivel máximo de RAID comúnmente utilizado en la actualidad, que es la protección de falla doble de RAID 6. Debido a que la confiabilidad del almacenamiento aumenta geoméricamente con esta cantidad de redundancia, la protección $+4n$ puede ser órdenes de magnitud más confiable que el RAID de hardware tradicional. Esta protección adicional significa que las unidades SATA de gran capacidad, como las unidades de 4 TB y 6 TB, se pueden agregar con confianza.

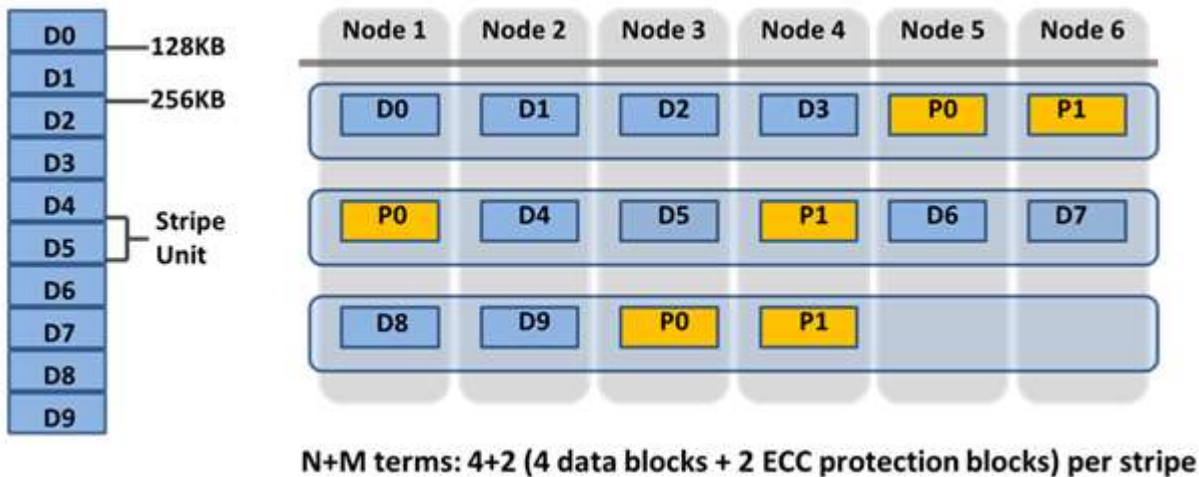


Figura 15: redundancia de OneFS: protección de código de eliminación N+M

Los clústeres más pequeños se pueden proteger con +1n, pero esto implica que, si bien se puede recuperar una sola unidad o un solo nodo, este no sería el caso para dos unidades en dos nodos diferentes. Las fallas de unidades son órdenes de magnitud más probables que las fallas de nodos. Para los clústeres con unidades de gran tamaño, es recomendable proporcionar protección contra fallas de unidades múltiples, a pesar de que se acepta la capacidad de recuperación de un solo nodo.

Para prever una situación en la que deseemos contar con redundancia de uno o dos discos, podemos crear grupos de protección de doble o triple ancho. Estos grupos de protección de doble o triple ancho “envolverán” una o dos veces el mismo conjunto de nodos a medida que se distribuyen. Dado que cada grupo de protección contiene exactamente dos discos de redundancia, este mecanismo permitirá que un clúster soporte dos o tres fallas de unidad o la falla de un nodo completo sin ninguna falta de disponibilidad de datos.

Y lo que es más importante para los clústeres pequeños, este método de fraccionado es altamente eficiente, con una eficiencia en disco de $M/(N+M)$. Por ejemplo, en un clúster de cinco nodos con protección de falla doble, si usáramos $N = 3$, $M = 2$, obtendríamos un grupo de protección 3+2 con una eficiencia de $1-2/5$ o 60 %. Con el mismo clúster de 5 nodos, pero con cada grupo de protección dispuesto en 2 fracciones, N sería 8 y $M = 2$, por lo tanto, podríamos obtener una eficiencia de $1-2/(8+2)$ u 80 % en disco, lo que mantiene nuestra protección contra fallas de dos unidades y sacrifica solo la protección contra fallas de dos nodos.

OneFS es compatible con varios esquemas de protección. Estos incluyen la protección universal +2d:1n, que brinda protección contra la falla de dos unidades o de un nodo.

① La práctica recomendada es usar el nivel de protección recomendado para una configuración de clúster específica. Este nivel recomendado de protección se marca claramente como “sugerido” en las páginas de configuración de los pools de almacenamiento de la interfaz de usuario web de OneFS y, por lo general, se configura de manera predeterminada. Para todas las configuraciones de hardware actuales de sexta generación, el nivel de protección recomendado es “+2d:1n”.

Los esquemas de protección híbrida son especialmente útiles para las configuraciones de nodos de alta densidad del chasis Gen6, en las que la probabilidad de que múltiples unidades fallen supera la probabilidad de que falle un nodo completo. En el caso improbable de que varios dispositivos tengan fallas simultáneas, por lo que el archivo está “más allá de su nivel de protección”, OneFS volverá a proteger todo lo posible e informará errores sobre los archivos individuales afectados en los registros del clúster.

OneFS también proporciona una variedad de opciones de espejeado que varían de 2x a 8x, lo que permite de dos a ocho espejeados del contenido especificado. Por ejemplo, los metadatos se espejean en un nivel por encima de FEC de manera predeterminada. Por ejemplo, si un archivo está protegido con +2n, su objeto de metadatos asociado se espejeará 3 veces.

La gama completa de niveles de protección de OneFS se resume en la siguiente tabla:

Nivel de protección	Descripción
+1n	Tolera la falla de 1 unidad O de 1 nodo
+2d:1n	Tolera la falla de 2 unidades O de 1 nodo
+2n	Tolera la falla de 2 unidades O de 2 nodos
+3d:1n	Tolera la falla de 3 unidades O de 1 nodo
+3d:1n1d	Tolera la falla de 3 unidades O de 1 nodo Y 1 unidad
+3n	Tolera la falla de 3 unidades o de 3 nodos
+4d:1n	Tolera la falla de 4 unidades o de 1 nodo
+4d:2n	Tolera la falla de 4 unidades o de 2 nodos
+4n	Tolera la falla de 4 nodos
De 2x a 8x	Espejeado de 2 a 8 nodos, según la configuración

OneFS permite que un administrador modifique la política de protección en tiempo real, mientras los clientes se conectan y leen y escriben datos.

① Tenga en cuenta que el aumento del nivel de protección de un clúster puede aumentar la cantidad de espacio consumido por los datos en el clúster.

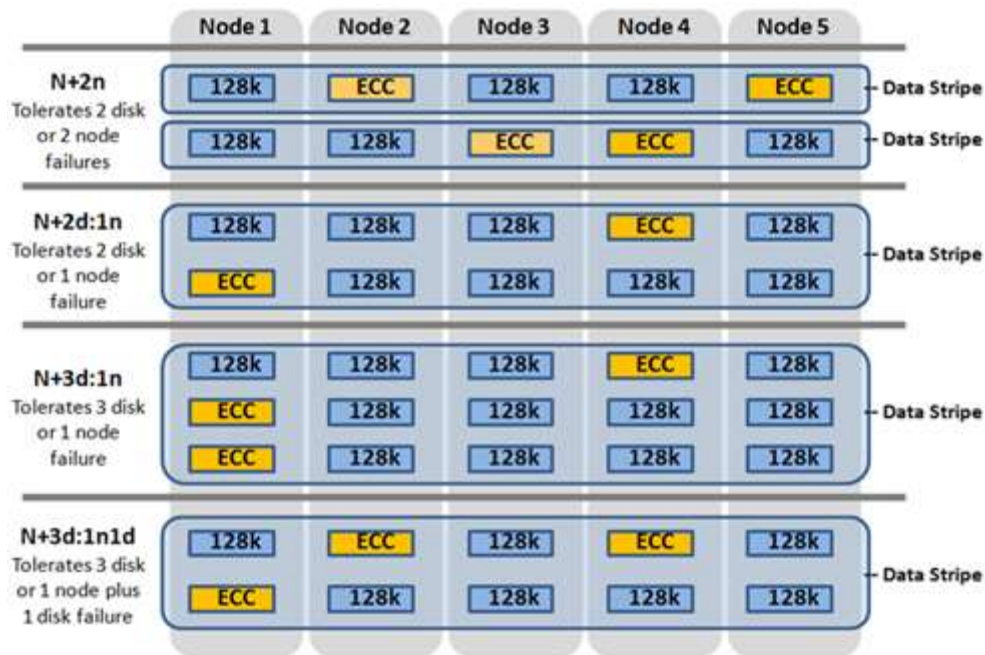


Figura 16: esquemas de protección de código de eliminación híbrida de OneFS

① OneFS también proporciona alertas de protección insuficiente para las instalaciones de clústeres nuevas. Si el clúster no está bien protegido, el sistema de registro de eventos del clúster (CELOG) genera alertas, advierte al administrador de la deficiencia de protección y recomienda un cambio al nivel de protección adecuado para la configuración específica de ese clúster.

📖 Encontrará más información disponible en la documentación técnica de [alta disponibilidad y protección de datos de OneFS](#).

Particionamiento automático

El marco de trabajo de SmartPools maneja la organización en niveles y la administración de los datos en OneFS. Desde un punto de vista de la eficiencia en el diseño y la protección de datos, SmartPools facilita la subdivisión de grandes cantidades de nodos homogéneos de alta capacidad en pools de discos más pequeños con un “tiempo promedio de pérdida de datos” (MTTDL) más amigable. Por ejemplo, un clúster H500 de 80 nodos se ejecutaría normalmente en un nivel de protección de un +3d:1n1d. Pero si se particiona en cuatro pools de veinte nodos, permitiría que cada pool se ejecutara con una protección +2d:1n, lo que reduce la sobrecarga de protección y mejora la utilización de espacio sin ningún aumento neto en la sobrecarga de administración.

A fin de mantener el objetivo de sencillez en la administración de almacenamiento, OneFS calculará y dividirá automáticamente el clúster en pools de discos o “pools de nodos”, los cuales están optimizados para el MTTDL y la utilización eficiente del espacio. Esto significa que las decisiones relacionadas con el nivel de protección, como el ejemplo del clúster de ochenta nodos, no dependen del cliente.

Con el aprovisionamiento automático, cada conjunto de hardware de nodo compatible se divide automáticamente en pools de discos que comprenden hasta cuarenta nodos y seis unidades por nodo. Estos pools de nodos cuentan de manera predeterminada con la protección +2d:1n, y varios pools pueden combinarse en niveles lógicos y administrarse con las políticas de pool de archivos de SmartPools. Mediante la subdivisión de los discos de un nodo en múltiples pools protegidos por separado, los nodos se vuelven significativamente más resistentes a múltiples fallas de discos de lo que anteriormente era posible.

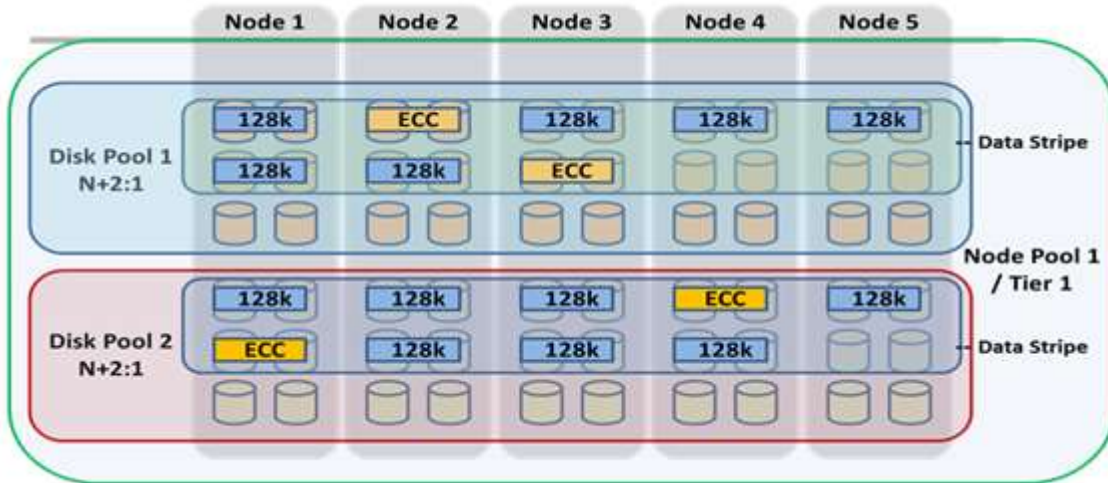


Figura 17: particionamiento automático con SmartPools

📖 Encontrará más información disponible en la [documentación técnica de SmartPools](#).

Las plataformas de hardware de PowerScale Gen6 cuentan con un diseño modular altamente denso que contiene cuatro nodos de PowerScale en un solo chasis de 4RU. Este enfoque mejora el concepto de pools de discos, pools de nodos y “vecindarios”, y agrega otro nivel de resiliencia al concepto de dominio de fallas de OneFS. Cada chasis de Gen6 contiene cuatro módulos de computación (uno por nodo) y cinco contenedores de unidades, o bahías, por nodo.

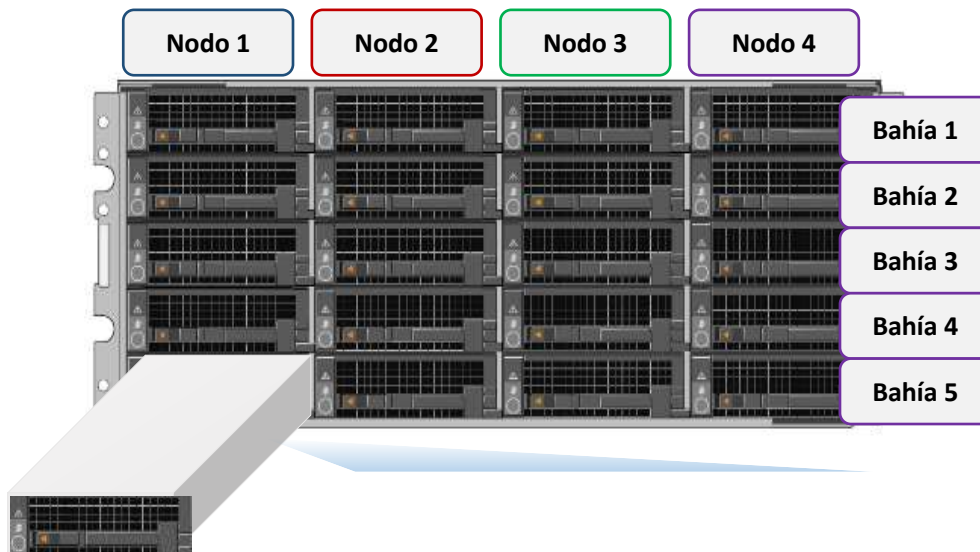


Figura 18. La vista frontal del chasis de plataforma Gen6 muestra las bahías de unidad.

Cada bahía es una bandeja que se desliza hacia la parte frontal del chasis y que contiene entre tres y seis unidades, según la configuración de un chasis específico. Los pools de discos son la unidad más pequeña dentro de la jerarquía de pools de almacenamiento. El aprovisionamiento de OneFS funciona sobre la premisa de dividir unidades de nodos similares en conjuntos, o pools de discos, donde cada pool representa un dominio de falla separado. Estos pools de discos están protegidos de manera predeterminada en +2d:1n (o la capacidad de resistir la falla de dos unidades o de un nodo completo).

Los pools de discos se distribuyen en las cinco bahías de cada nodo de Gen6. Por ejemplo, un nodo con tres unidades por bahía tendrá la siguiente configuración de pool de discos:

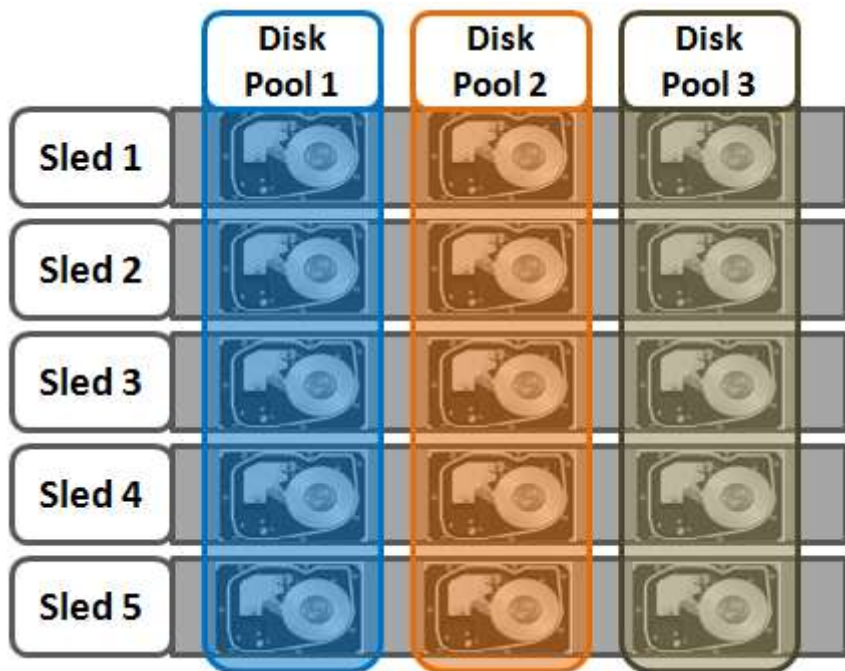


Figura 19. Pools de discos OneFS

Los pools de nodos son grupos de pools de discos que se distribuyen entre nodos de almacenamiento similares (clases de compatibilidad). Esto se muestra en la figura 20 que aparece a continuación. Varios grupos de diferentes tipos de nodos pueden trabajar en conjunto en un clúster único y heterogéneo. Por ejemplo: un pool de nodos de la serie F para las aplicaciones con gran actividad de I/O, un pool de nodos de la serie H que se utiliza principalmente para cargas de trabajo secuenciales y de alta simultaneidad, y un pool de nodos de la serie A que se utiliza principalmente para cargas de trabajo nearline o de archiving profundo.

Esto permite que OneFS presente un solo pool de recursos de almacenamiento que consta de varios tipos de medios de unidad: SSD, SAS de alta velocidad, SATA de gran capacidad, etc., lo que proporciona una variedad de características distintas de rendimiento, protección y capacidad. A su vez, este pool de almacenamiento heterogéneo puede admitir una amplia variedad de aplicaciones y requisitos de cargas de trabajo con un único punto de administración unificado. También facilita la combinación de hardware más antiguo y más reciente, lo que permite una protección simple de la inversión incluso a través de las distintas generaciones de productos y actualizaciones de hardware transparentes.

Cada pool de nodos contiene solamente pools de discos del mismo tipo de nodos de almacenamiento, y un pool de discos puede pertenecer a exactamente un pool de nodos. Por ejemplo, los nodos de la serie F con unidades SSD de 1,6 TB estarían en un pool de nodos, mientras que los nodos de la serie A con unidades SATA de 10 TB estarían en otro. En la actualidad, se requiere un mínimo de 4 nodos (un chasis) por pool de nodos para el hardware Gen6, como PowerScale H700, o tres nodos por pool para nodos autónomos como PowerScale F900.

Los “vecindarios” de OneFS son dominios de fallas dentro de un pool de nodos, y su propósito es mejorar la confiabilidad en general y brindar protección contra la falta de disponibilidad de datos debido a la eliminación accidental de la bahías de unidad. Para nodos autocontenidos como PowerScale F200, OneFS tiene un tamaño ideal de 20 nodos por pool de nodos y un tamaño máximo de 39 nodos. Tras la adición del 40.º nodo, los nodos se dividen en dos vecindarios de veinte nodos.

Con la plataforma Gen6, el tamaño ideal de un vecindario cambia de 20 a 10 nodos. Esto brindará protección contra las fallas simultáneas del registro de pares de nodos y las fallas de chasis completo.

Los nodos compañeros son nodos cuyos registros están espejados. Con la plataforma Gen6, en lugar de que cada nodo almacene su registro en NVRAM como en las plataformas anteriores, los registros de los nodos se almacenan en unidades SSD, y cada registro tiene una copia espejada en otro nodo. El nodo que contiene el registro espejado se conoce como el nodo compañero. Se obtienen varios beneficios de confiabilidad de los cambios en el registro. Por ejemplo, las unidades SSD son más persistentes y confiables que la NVRAM, lo cual requiere una batería cargada para conservar el estado. Además, con el registro espejado, ambas unidades de registro deben dejar de funcionar antes de que un registro se considere perdido. Por lo tanto, a menos que ambas unidades de registro espejadas fallen, los dos nodos compañeros pueden funcionar de manera normal.

Con la protección de nodos compañeros, los nodos se colocarán en diferentes vecindarios cuando sea posible y, por lo tanto, en diferentes dominios de fallas. La protección de nodos compañeros es posible una vez que el clúster alcanza cinco chasis completos (20 nodos) cuando, después de la división del primer vecindario, OneFS coloca los nodos compañeros en diferentes vecindarios:

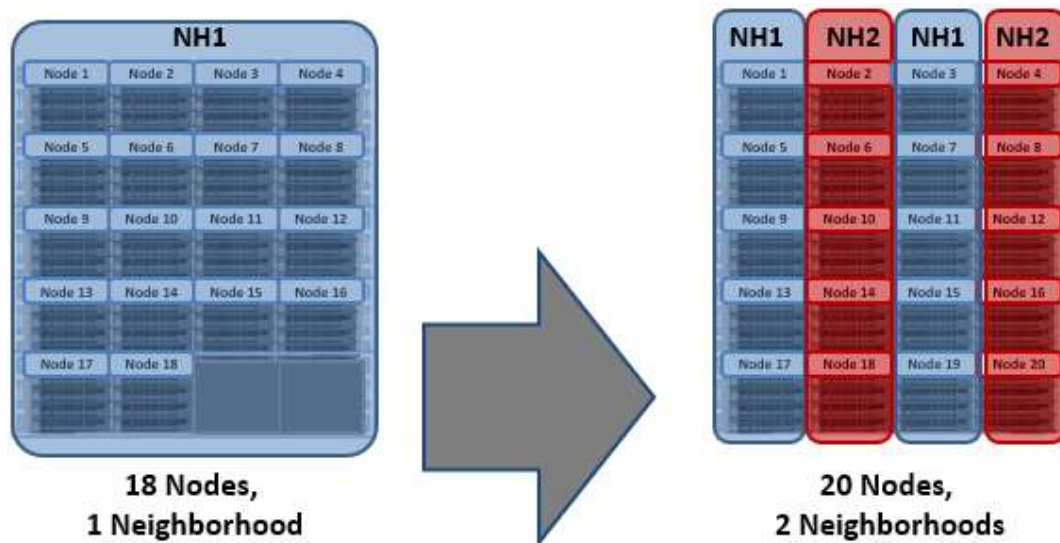


Figura 20. División a dos vecindarios de veinte nodos

La protección de nodos compañeros aumenta la confiabilidad, ya que si ambos nodos fallan, se encontrarán en distintos dominios de fallas, por lo que sus dominios solamente sufrirán la pérdida de un nodo.

Con la protección del chasis, cada uno de los cuatro nodos dentro de un chasis se colocará en un vecindario diferente cuando sea posible. La protección del chasis se vuelve posible con 40 nodos, ya que si el vecindario se divide en 40 nodos, permite que cada nodo de un chasis se coloque en un vecindario diferente. Por lo tanto, cuando un clúster Gen6 de 38 nodos se expande a 40 nodos, los dos vecindarios existentes se dividirán en cuatro vecindarios de 10 nodos:

La protección del chasis garantiza que, si se produce un error en un chasis completo, cada dominio de falla solo perderá un nodo.

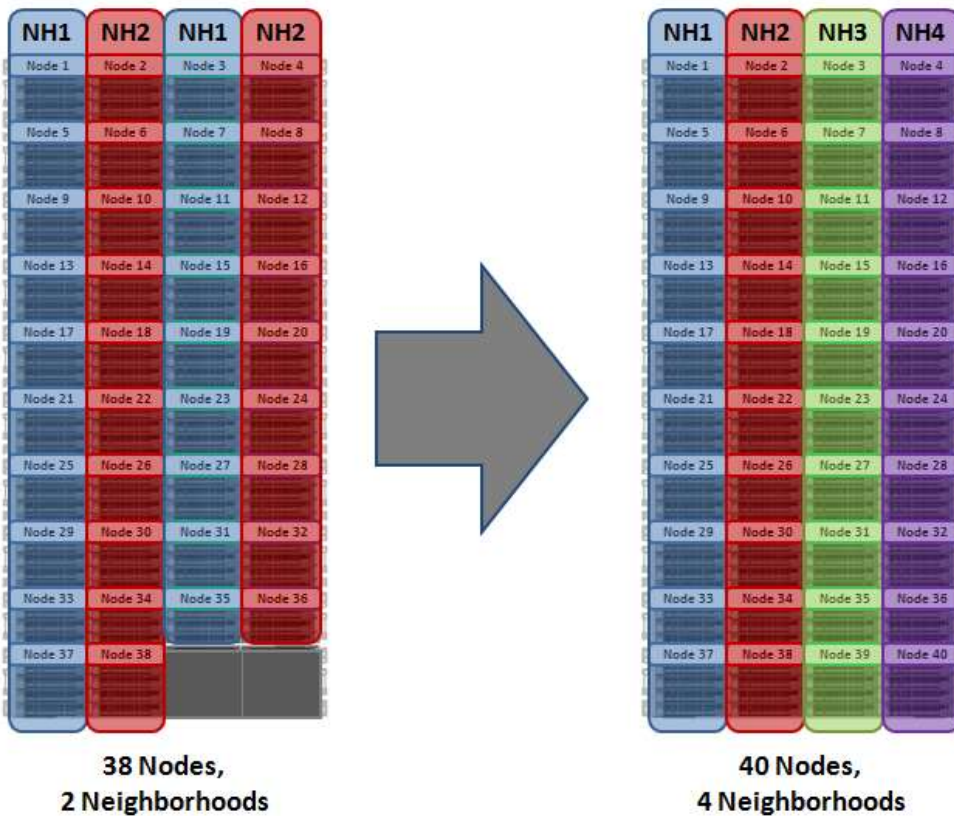


Figura 21. Vecindarios de OneFS: división de cuatro vecindarios

❶ Un clúster de 40 nodos o de mayor tamaño con cuatro vecindarios, protegidos con el nivel predeterminado de +2d:1n, puede admitir la falla de un nodo único por cada vecindario. Esto protege el clúster frente a la falla de un solo chasis Gen6.

En general, un clúster de la plataforma Gen6 tendrá una confiabilidad de al menos un orden de magnitud superior al de los clústeres de la generación anterior con una capacidad similar como resultado directo de las siguientes mejoras:

- Registros espejados
- Vecindarios más pequeños
- Unidades de arranque espejadas

Compatibilidad


Algunos tipos de nodos similares, pero no idénticos, se pueden aprovisionar a un pool de nodos existente mediante la compatibilidad de nodos. OneFS requiere que un pool de nodos contenga un mínimo de tres nodos.

❶ Debido a diferencias arquitectónicas significativas, no hay compatibilidades de nodos entre el plataforma Gen6, las generaciones de hardware anteriores o los nodos PowerScale.

OneFS también contiene una opción de compatibilidad de unidades SSD, lo que permite que los nodos con unidades SSD que presentan una capacidad distinta se aprovisionen en un solo pool de nodos.

La compatibilidad SSD se crea y se describe en la lista OneFS WebUI SmartPools Compatibilities, y también se muestra en la lista Tiers & Node Pools.

❶ Cuando se crea esta compatibilidad de unidades SSD, OneFS comprueba automáticamente que los dos pools que se fusionarán tengan la mismas características de cantidad de unidades SSD, niveles, protección solicitada y configuración de caché L3. Si estos ajustes difieren, la interfaz de usuario web de OneFS solicitará la consolidación y la alineación de estos ajustes.

 Encontrará más información disponible en la [documentación técnica de SmartPools](#).

Protocolos compatibles

Los clientes con credenciales y privilegios adecuados pueden crear, modificar y leer datos con uno de los métodos estándares admitidos para la comunicación con el clúster:

- NFS (sistema de archivos de red)
- SMB/CIFS (bloque de mensajes del servidor/sistema de archivos común de Internet)
- Protocolo de transferencia de archivos (FTP)
- HTTP (protocolo de transferencia de hipertexto)
- HDFS (sistema de archivos distribuido Hadoop)
- API REST (interfaz de programación de aplicaciones de transferencia de estado representacional)
- S3 (API de almacenamiento de objetos)


Para el protocolo NFS, OneFS es compatible con NFSv3 y NFSv4, además de NFSv4.1 en OneFS 9.3. Además, OneFS 9.2 y versiones posteriores incluyen compatibilidad con NFSv3overRDMA.

Del lado de Microsoft Windows, el protocolo SMB admite hasta la versión 3. Como parte del dialecto de SMB3, OneFS admite las siguientes funciones:

- Múltiples rutas de SMB3
- Testigo y disponibilidad continua de SMB3
- Cifrado de SMB3

El cifrado de SMB3 se puede configurar por recurso compartido, por zona o en todo el clúster. Solo los sistemas operativos que admiten el cifrado de SMB3 pueden trabajar con recursos compartidos cifrados. Estos sistemas operativos también pueden trabajar con recursos compartidos no cifrados si el clúster está configurado para permitir conexiones no cifradas. Estos sistemas operativos también pueden trabajar con recursos compartidos no cifrados solo si el clúster está configurado para permitir conexiones no cifradas.

La raíz del sistema de archivos para todos los datos en el clúster es /ifs (el sistema de archivos OneFS). Esto se puede presentar a través del protocolo SMB como un recurso compartido "ifs" (\\<cluster_name>\ifs) y a través del protocolo NFS como una exportación "/ifs" (<cluster_name>:/ifs).

 Los datos son comunes entre todos los protocolos, por lo que los cambios realizados en el contenido del archivo a través de un protocolo de acceso se pueden ver al instante desde los demás protocolos.

OneFS ofrece soporte completo para los entornos IPv4 e IPv6 en las redes Ethernet de front-end, SmartConnect y la variedad completa de protocolos de almacenamiento y herramientas de administración.

Además, OneFS CloudPools es compatible con las API de almacenamiento de los siguientes proveedores de nube, lo que permite que los archivos se conviertan en stubs en una gran cantidad de destinos de almacenamiento, incluidos los siguientes:

- Amazon Web Services S3
- Microsoft Azure
- Google Cloud Service
- Alibaba Cloud
- Dell EMC ECS
- OneFS RAN (acceso de RESTful al espacio de nombres)

 Encontrará más información disponible en la [Guía de administración de CloudPools](#).

Operaciones no disruptivas: compatibilidad con protocolos

OneFS contribuye a la disponibilidad de datos mediante la compatibilidad con la conmutación por error y la conmutación por recuperación dinámicas de NFSv3 y NFSv4 para clientes Linux y UNIX, y la disponibilidad continua de SMB3 para clientes Windows. Esto garantiza que, cuando se produce la falla de un nodo u ocurre un mantenimiento preventivo, todas las lecturas y las escrituras en transferencia se trasladan a otro nodo del clúster para completar su operación sin ninguna interrupción para los usuarios o las aplicaciones.

Durante una falla, los clientes se distribuyen de manera equitativa entre todos los nodos restantes del clúster, garantizando un impacto mínimo en el rendimiento. Si un nodo deja de funcionar por cualquier motivo, incluida una falla, las direcciones IP virtuales de ese nodo se migran sin problemas a otro nodo del clúster.

Cuando el nodo fuera de línea vuelve a ponerse en línea, SmartConnect rebalancea automáticamente los clientes NFS y SMB3 en todo el clúster para garantizar el máximo nivel de utilización de almacenamiento y rendimiento. En el caso del mantenimiento periódico del sistema y las actualizaciones de software, esta funcionalidad permite realizar actualizaciones graduales por nodo, lo cual proporciona una disponibilidad completa durante la ventana de mantenimiento.

Filtrado de archivo

El filtrado de archivos de OneFS se puede usar en los clientes NFS y SMB para permitir o impedir las escrituras en una zona de exportación, uso compartido o acceso. Esta función impide el bloqueo de ciertos tipos de extensiones de archivos para los archivos que podrían causar problemas de seguridad, las interrupciones en la productividad, los problemas de rendimiento o el desorden de almacenamiento. La configuración se puede realizar a través de una lista de exclusión, la cual bloquea las extensiones de archivo explícitas, o una lista de inclusión, que permite explícitamente las escrituras de ciertos tipos de archivo solamente.

Desduplicación de datos: SmartDedupe

El producto SmartDedupe maximiza la eficiencia del almacenamiento de un clúster mediante la reducción de la cantidad de almacenamiento físico que se necesita para guardar los datos de la organización. La eficiencia se logra escaneando los datos en disco en busca de bloques idénticos, para luego eliminar los duplicados. Por lo general, este enfoque se conoce como desduplicación posterior al proceso o asíncrona.

Después de que se descubren los bloques duplicados, SmartDedupe transfiere una sola copia de esos bloques a un conjunto especial de archivos conocidos como almacenes ocultos. Durante este proceso, los bloques duplicados se eliminan de los archivos reales y se reemplazan por punteros hacia los almacenes ocultos.

Con la desduplicación posterior al proceso, los datos nuevos se almacenan primero en el dispositivo de almacenamiento. Después, un proceso subsiguiente analiza los datos en busca de elementos comunes. Esto significa que el rendimiento inicial de escritura o modificación de archivos no se ve afectado, ya que no se requiere ninguna computación adicional en la ruta de escritura.

Arquitectura de SmartDedupe

La arquitectura de OneFS SmartDedupe consta de cinco módulos básicos:

- Ruta de control de desduplicación
- Trabajo de desduplicación
- Motor de desduplicación
- Almacén oculto
- Infraestructura de desduplicación

La ruta de control de SmartDedupe consta de la interfaz de administración web (WebUI) de OneFS, la interfaz de la línea de comandos (CLI) y la API de la plataforma RESTful, y es responsable de administrar la configuración, la programación y el control del trabajo de desduplicación. El trabajo en sí es un proceso en segundo plano altamente distribuido que administra la orquestación de la desduplicación en todos los nodos del clúster. El control del trabajo abarca el escaneo del sistema de archivos, la detección y el uso compartido de bloques de datos coincidentes en conjunto con el motor de desduplicación. La capa de infraestructura de desduplicación es el módulo de kernel que lleva a cabo la consolidación de bloques de datos compartidos en almacenes ocultos, los contenedores del sistema de archivos que almacenan los bloques de datos físicos y las referencias, o los punteros, a bloques compartidos. Estos elementos se describen más detalladamente a continuación.



Figura 22: arquitectura modular de OneFS SmartDeduplication

📖 Encontrará más información disponible en la documentación técnica de [OneFS SmartDeduplication](#).

Áreas de almacenamiento ocultas

Los almacenes ocultos de OneFS son contenedores del sistema de archivos que permiten que los datos se almacenen de manera compartida. Por lo tanto, los archivos en OneFS pueden contener datos físicos y punteros, o referencias, a bloques compartidos en almacenes ocultos.

Los almacenes ocultos son similares a los archivos normales, pero no suelen contener todos los metadatos que generalmente se asocian a inodos de archivos regulares. En especial, los atributos basados en tiempo (hora de creación, hora de modificación, etc.) no se mantienen explícitamente. Cada almacén oculto puede contener hasta 256 bloques, con 32 000 archivos que pueden hacer referencia a cada bloque. Si se supera este límite de referencia de 32 000, se crea un nuevo almacén oculto. Además, los almacenes ocultos no hacen referencia a otros almacenes ocultos. Y las instantáneas de los almacenes ocultos no están permitidas, ya que no tienen enlaces físicos.

📌 Los almacenes ocultos también se utilizan para los clones de archivos de OneFS y la eficiencia del almacenamiento de archivos pequeños (SFSE), además de la deduplicación.

Eficiencia del almacenamiento de archivos pequeños

Otro consumidor principal de almacenes ocultos es la eficiencia del almacenamiento de archivos pequeños de OneFS. Esta función maximiza la utilización de espacio de un clúster mediante la reducción de la cantidad de almacenamiento físico necesario para alojar archivos pequeños que a menudo conforman un conjunto de datos de archivado, como se puede observar en los flujos de trabajo de PACS de los servicios de salud.

La eficiencia se logra mediante el escaneo de los datos en disco para buscar archivos pequeños, que están protegidos por espejados de copia completa, y su empaquetamiento en almacenes ocultos. A continuación, estos almacenes ocultos se protegen con paridad, en lugar de espejado, y suelen proporcionar una eficiencia del almacenamiento del 80 % o más.

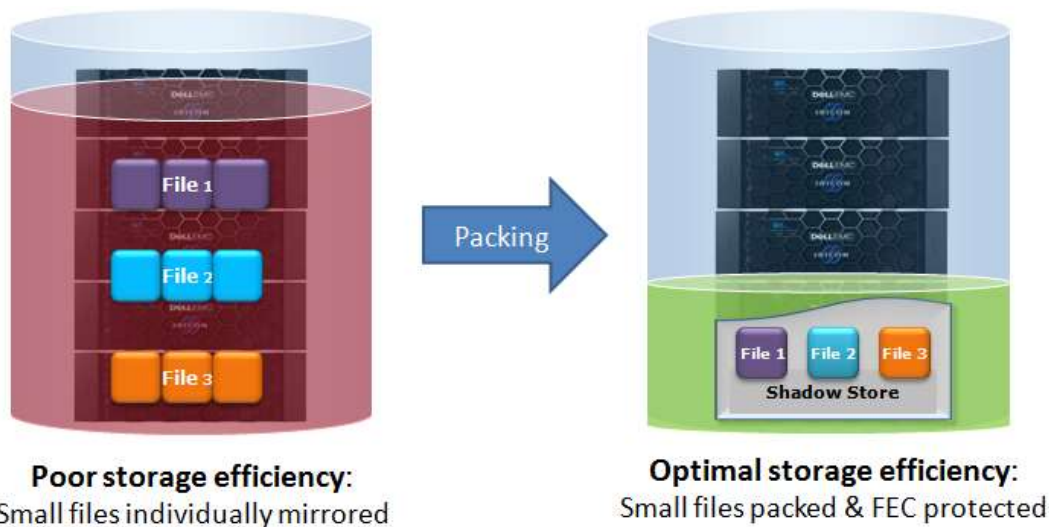


Figura 23: contenedores de archivos pequeños

La eficiencia del almacenamiento de archivos pequeños traza una pequeña pérdida de rendimiento de latencia por una mejor utilización del almacenamiento. Los archivos archivados obviamente mantienen su capacidad de escritura, pero cuando se eliminan, se truncan o se sobrescriben los archivos en contenedores con referencias ocultas, se pueden dejar bloques no referenciados en almacenes ocultos. Posteriormente, estos bloques se liberan y pueden generar brechas que reducen la eficiencia del almacenamiento.

La pérdida de eficiencia real depende del diseño de nivel de protección utilizado por el almacén oculto. Los grupos de protección más pequeños son más susceptibles, al igual que los archivos en contenedores, ya que todos los bloques en contenedores tienen al menos un archivo de referencia y los tamaños empaquetados (tamaño de archivo) son pequeños.

Se proporciona un desfragmentador para reducir la fragmentación de los archivos como resultado de sobrescrituras y eliminaciones. Este desfragmentador de almacenes ocultos está integrado en el trabajo de ShadowStoreDelete. El proceso de desfragmentación funciona mediante la división de cada archivo en contenedores en fragmentos lógicos (aprox. 32 MB cada uno) y la evaluación de cada fragmento para la fragmentación.

Si la eficiencia del almacenamiento de un fragmento fragmentado está por debajo del objetivo, ese fragmento se procesa mediante la evacuación de los datos a otra ubicación. La eficiencia predeterminada de destino es del 90 % de la eficiencia del almacenamiento máxima disponible en el nivel de protección que utiliza el almacén oculto. Los tamaños de grupos de protección más grandes pueden tolerar un nivel más alto de fragmentación antes de que la eficiencia del almacenamiento disminuya por debajo de este umbral.

Reducción de datos en línea

La reducción de datos en línea de OneFS está disponible en los nodos todo flash F900, F810, F600 y F200, en los chasis híbridos H700/7000 y H5600, y en la plataforma de archivado A300/3000. La arquitectura de OneFS consta de los siguientes componentes principales:

- Plataforma de reducción de datos
- Motor de compresión y mapa de fragmentos
- Fase de eliminación con bloques de ceros
- Índice de deduplicación en la memoria e infraestructura de un área de almacenamiento oculta
- Infraestructura de generación de informes y alertas de reducción de datos
- Ruta de control de reducción de datos

La ruta de escritura de reducción de datos en línea consta de tres fases principales:

- Eliminación con bloques de ceros
- Deduplicación en línea
- Compresión en línea

Si la compresión y la deduplicación en línea están habilitadas en un clúster, la eliminación con bloques de ceros se realiza primero, seguido de la deduplicación y, a continuación, la compresión. Este orden permite que cada fase reduzca el alcance de trabajo en cada fase subsiguiente.



Figura 24: flujo de trabajo de reducción de datos en línea.

El F810 incluye una funcionalidad de descarga de compresión de hardware, con cada nodo en un chasis de F810 que contiene un adaptador Mellanox Innova-2 Flex. Esto significa que el adaptador de Mellanox lleva a cabo la compresión y la descompresión de manera transparente con una latencia mínima, lo que evita la necesidad de consumir recursos de CPU y memoria costosos de un nodo.

El motor de compresión de hardware de OneFS utiliza zlib, con una implementación de software de igzip para los nodos PowerScale F900, F810, F600, F200, H700/7000, H5600 y A300/3000. La compresión de software también se utiliza como reserva en caso de una falla de hardware de compresión, y en un clúster mixto, para su uso en nodos que no son de F810 sin una funcionalidad de compresión de hardware y como reserva en caso de una falla de hardware de compresión. OneFS emplea un tamaño de fragmento de compresión de 128 KB, con cada fragmento compuesto por 16 bloques de datos de 8 KB. Esto es ideal, ya que también tiene el mismo tamaño que OneFS utiliza para sus unidades de fracción de protección de datos, lo que proporciona sencillez y eficiencia debido a que se evita la sobrecarga de paquetes de fragmentos adicionales.

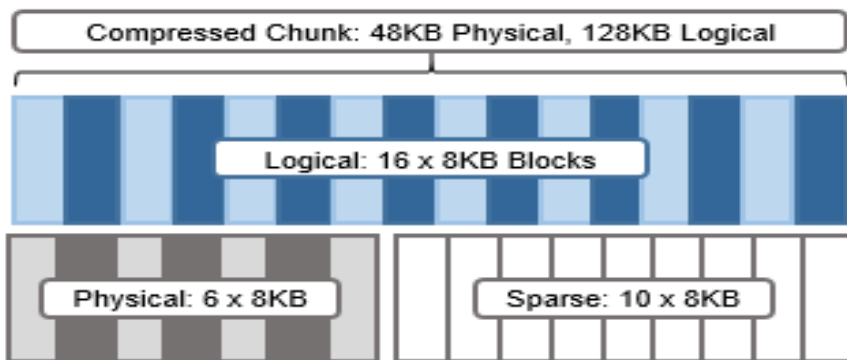


Figura 25: fragmentos de compresión y superposición transparente de OneFS.

Considere el diagrama anterior. Después de la compresión, este fragmento reduce su tamaño de 16 bloques a 6 bloques de 8 KB. Esto significa que este fragmento ahora tiene un tamaño físico de 48 KB. OneFS proporciona una superposición lógica transparente a los atributos físicos. Esta superposición describe si los datos de respaldo están comprimidos o no y qué bloques del fragmento son físicos o dispersos, de manera que los consumidores del sistema de archivos no se vean afectados por la compresión. Por lo tanto, el fragmento comprimido se representa de manera lógica con un tamaño de 128 KB, independientemente de su tamaño físico real.

El ahorro de eficiencia debe ser de al menos 8 KB (un bloque) a fin de que se produzca la compresión; de lo contrario, ese fragmento o archivo se transferirá y permanecerá en su estado original sin comprimir. Por ejemplo, un archivo de 16 KB que genera 8 KB (un bloque) de ahorro se podría comprimir. Una vez que un archivo termina de comprimirse, se protege mediante FEC.

Los fragmentos de compresión nunca atravesarán pools de nodos. Esto evita la necesidad de descomprimir o recomprimir datos para cambiar los niveles de protección, realizar escrituras recuperadas o cambiar los límites de los grupos de protección.

Escalamiento dinámico/según demanda

Rendimiento y capacidad

En contraste con los sistemas de almacenamiento tradicionales que deben realizar un “escalamiento vertical” cuando se requiere rendimiento o capacidad adicional, OneFS permite que un sistema de almacenamiento realice un “escalamiento horizontal” para aumentar de manera transparente el sistema de archivos o el volumen existente a petabytes de capacidad, a la vez que aumenta el rendimiento en conjunto de manera lineal.

Agregar funcionalidades de capacidad y rendimiento a un clúster es mucho más fácil que con otros sistemas de almacenamiento, ya que requiere tres sencillos pasos para el administrador de almacenamiento: agregar otro nodo al rack, conectar el nodo a la red de back-end e indicarle al clúster que agregue el nodo adicional. El nuevo nodo ofrece capacidad y rendimiento adicionales, ya que cada nodo incluye rutas de control de CPU, memoria, caché, red, NVRAM e I/O.

La función Autobalance de OneFS transferirá automáticamente los datos en la red de back-end de manera automática y coherente, de modo que los datos existentes que residen en el clúster puedan transferirse a este nuevo nodo de almacenamiento. Este rebalanceo automático garantiza que el nodo nuevo no se convierta en un punto problemático para los datos nuevos y que esos datos existentes puedan obtener los beneficios de un sistema de almacenamiento más eficiente. La función Autobalance de OneFS es también completamente transparente para el usuario final y se puede ajustar para minimizar el impacto en las cargas de trabajo de alto rendimiento. Esta función solamente permite a OneFS escalar de manera transparente y dinámica desde TB a PB sin agregar más tiempo de administración para el administrador, ni incrementar la complejidad dentro del sistema de almacenamiento.

Un sistema de almacenamiento a gran escala debe proporcionar el rendimiento requerido para una variedad de flujos de trabajo, sean estos secuenciales, simultáneos o aleatorios. Habrá distintos flujos de trabajo entre aplicaciones y dentro de cada aplicación. OneFS satisface todas estas necesidades simultáneamente con software inteligente. Lo más importante es que con OneFS, el rendimiento y los IOPS escalan de manera lineal con la cantidad de nodos presentes en un solo sistema. Debido a la distribución de datos balanceada, el rebalanceo automático y el procesamiento distribuido, OneFS puede aprovechar CPU, puertos de red y memoria adicionales a medida que el sistema escala.

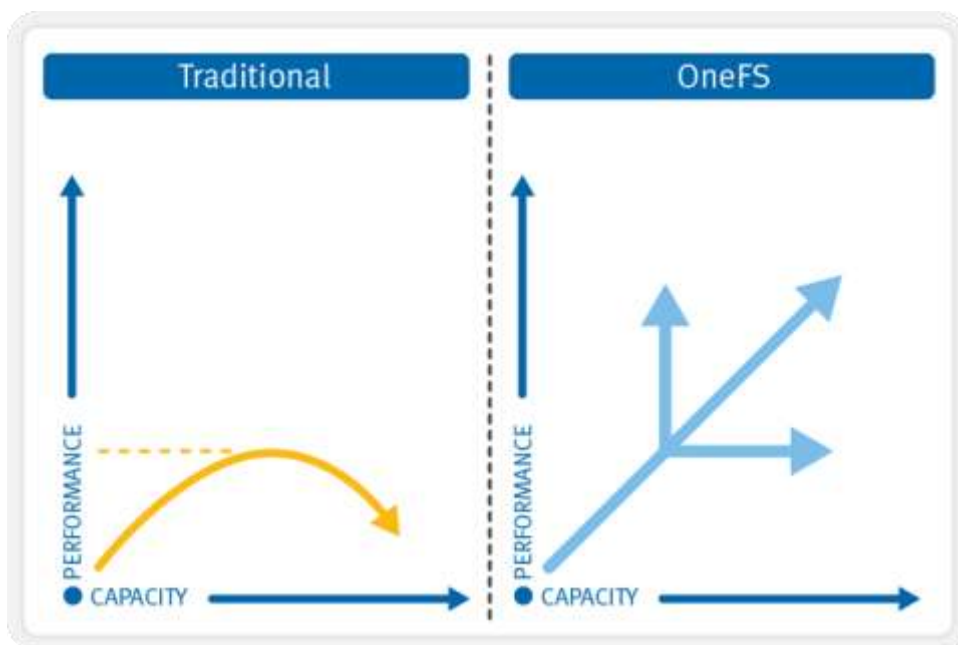


Figura 26: escalabilidad lineal de OneFS

Interfaces

Los administradores pueden usar varias interfaces para administrar un clúster de almacenamiento en sus entornos:

- Interfaz de usuario de administración web (“WebUI”)
- Interfaz de la línea de comandos a través del acceso de red SSH o la conexión en serie RS232
- Panel LCD en los propios nodos para las funciones simples de adición/eliminación
- API de plataforma RESTful para el control programático y la automatización de la configuración y la administración de clústeres.

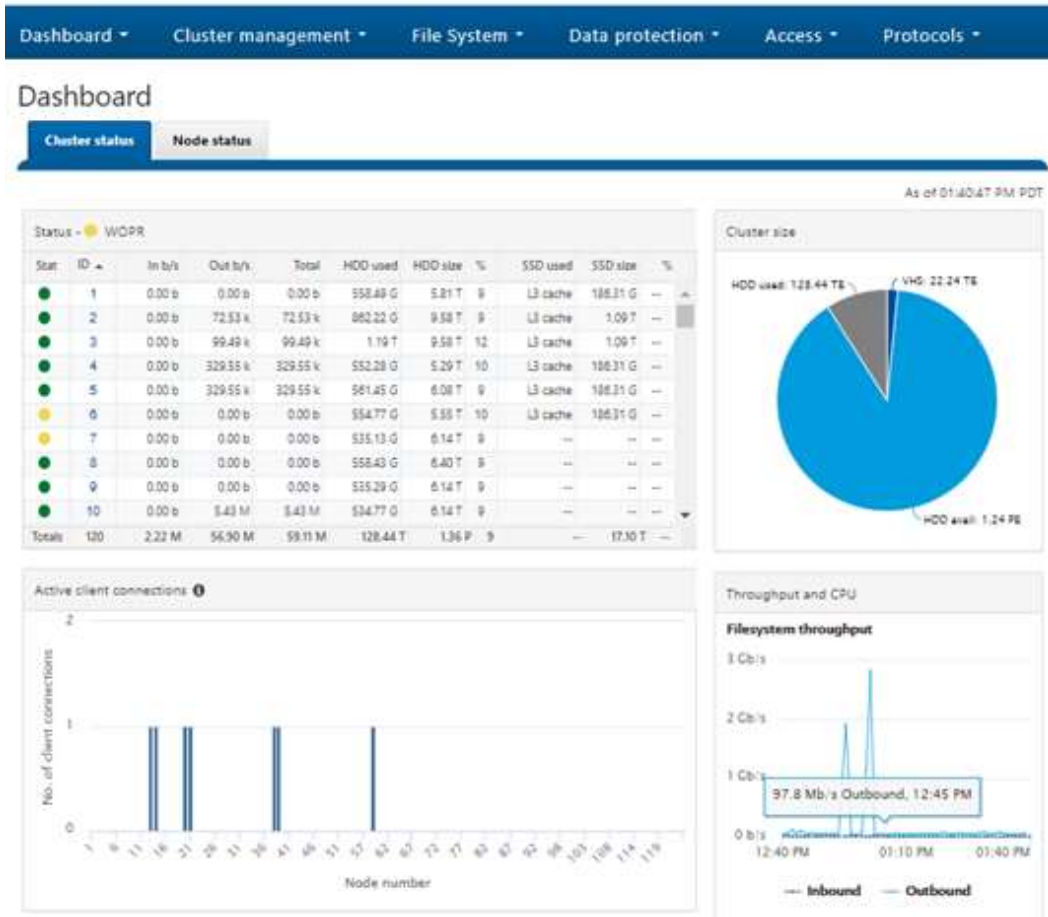


Figura 27: interfaz de usuario web de OneFS

Encontrará más información sobre los comandos y la configuración de funciones de OneFS en la [Guía de administración de OneFS](#).

Autenticación y control de acceso

Los servicios de autenticación ofrecen una capa de seguridad, ya que verifican las credenciales de los usuarios antes de permitirles acceder a los archivos y modificarlos. OneFS admite cuatro métodos para la autenticación de usuarios:

- Active Directory (AD)
- LDAP (protocolo ligero de acceso a directorios)
- NIS (servicio de información de red)
- Usuarios y grupos locales

OneFS admite el uso de más de un tipo de autenticación. Sin embargo, se recomienda comprender completamente las interacciones entre los tipos de autenticación antes de habilitar varios métodos en el clúster. Consulte la documentación del producto para obtener información detallada sobre cómo configurar correctamente varios modos de autenticación.

Active Directory

Active Directory, una implementación de LDAP de Microsoft, es un servicio de directorios que puede almacenar información sobre los recursos de red. Si bien Active Directory puede desempeñar muchas funciones, el motivo principal para vincular el clúster a un dominio es la ejecución de la autenticación de grupos y usuarios.

Puede configurar y administrar la configuración de Active Directory de un clúster desde la interfaz de administración web o desde la interfaz de la línea de comandos; sin embargo, se recomienda utilizar la administración web siempre que sea posible.

Cada nodo del clúster comparte la misma cuenta de máquina de Active Directory, lo que facilita en gran medida la administración.

LDAP

El protocolo ligero de acceso a directorios (LDAP) es un protocolo de red que le permite definir, consultar y modificar recursos y servicios. Una de las principales ventajas de LDAP es la naturaleza abierta de sus servicios de directorio y la capacidad de utilizar LDAP en varias plataformas. El sistema de almacenamiento en clúster puede utilizar LDAP para autenticar usuarios y grupos a fin de otorgarles acceso al clúster.

NIS

El Network Information Service (NIS), diseñado por Sun Microsystems, es un protocolo de servicios de directorio que OneFS puede usar para autenticar usuarios y grupos cuando se accede al clúster. NIS, que a veces se denomina Páginas amarillas (YP), es diferente de NIS+, el cual no es compatible con OneFS.

Usuarios locales

OneFS es compatible con la autenticación de usuarios y grupos locales. Puede crear cuentas de usuarios y grupos locales directamente en el clúster mediante la interfaz de usuario web. La autenticación local es útil cuando los servicios de directorio (Active Directory, LDAP o NIS) no se están usando, o cuando una aplicación o un usuario específicos necesitan acceder al clúster.

Zonas de acceso

Las zonas de acceso ofrecen un método para particionar el acceso a un clúster de forma lógica y asignar recursos a unidades autónomas, lo que proporciona un entorno de grupo de usuarios, o multiusuario, compartido. Para facilitar esto, las zonas de acceso unen los tres componentes principales de acceso externo:

- Configuración de redes del clúster
- Acceso multiprotocolo a archivos
- Autenticación


Por lo tanto, las zonas de SmartConnect están asociadas con un conjunto de recursos compartidos SMB, exportaciones de NFS, racks HDFS y uno o más proveedores de autenticación por zona para el control de acceso. Esto brinda los beneficios de un solo sistema de archivos administrado de forma centralizada que se puede aprovisionar y proteger para varios grupos de usuarios. Esto es especialmente útil para los entornos empresariales donde un departamento de TI central ofrece servicios a múltiples unidades de negocio independientes. Se puede observar otro ejemplo durante una iniciativa de consolidación de servidores, cuando se combinan varios servidores de archivos de Windows que están unidos a bosques de Active Directory independientes y no confiables.

Con las zonas de acceso, la zona de acceso del sistema incorporada incluye una instancia de cada proveedor de autenticación compatible, todos los recursos compartidos de SMB disponibles y todas las exportaciones de NFS disponibles de manera predeterminada.

Estos proveedores de autenticación pueden incluir varias instancias de Microsoft Active Directory, LDAP, NIS y bases de datos de grupos o usuarios locales.

Administración basada en funciones

La administración basada en funciones es un sistema de control de acceso basado en funciones (RBAC) de administración de clústeres que divide los poderes de los usuarios “raíz” y “administrador” en privilegios más granulares y permite la asignación de estos a funciones específicas. Estas funciones se pueden otorgar a otros usuarios que no tienen privilegios. Por ejemplo, es posible asignarle al personal de operaciones del centro de datos derechos de solo lectura para todo el clúster, lo que permite el acceso de monitoreo completo, pero sin ningún cambio en la configuración. OneFS proporciona una recopilación de funciones incorporadas, incluido el administrador de auditorías, sistemas y seguridad, además de la capacidad de crear funciones definidas personalizadas, ya sea por zona de acceso o en el clúster. La administración basada en funciones está integrada con la interfaz de la línea de comandos, la interfaz de usuario web y la API de plataforma de OneFS.


 Para obtener más información sobre la administración de identidades, la autenticación y el control de acceso en entornos de protocolos múltiples, consulte la [Guía de seguridad multiprotocolo de OneFS](#).

Auditoría de OneFS

OneFS proporciona la capacidad de auditar la configuración del sistema y la actividad de protocolo NFS, SMB y HDFS en un clúster. Esto permite que las organizaciones cumplan con diversas exigencias de gobierno corporativo de datos y cumplimiento de normas a las que pueden estar vinculadas.

Todos los datos de auditoría están almacenados y protegidos en el sistema de archivos del clúster y organizados según temas de auditoría. Desde aquí, los datos de auditoría se pueden exportar a través del marco de trabajo de Dell EMC Common Event Enabler (CEE) a aplicaciones de otros fabricantes, como Varonis DatAdvantage y Symantec Data Insight. La auditoría del protocolo OneFS se puede habilitar por zona de acceso, lo que permite un control granular en todo el clúster.

Un clúster puede escribir eventos de auditoría en un máximo de cinco servidores CEE por nodo en una configuración paralela con balanceo de carga. Esto permite que OneFS proporcione una solución de auditoría de punto a punto de nivel empresarial.

 Encontrará más información disponible en la documentación técnica de las [auditorías de OneFS](#).

Actualización de software

La actualización a la versión más reciente de OneFS le permite aprovechar todas las funciones, las reparaciones y las funcionalidades nuevas. Los clústeres se pueden actualizar mediante dos métodos: actualización simultánea o gradual

Actualización simultánea

Una actualización simultánea instala el nuevo sistema operativo y reinicia todos los nodos del clúster al mismo tiempo. Una actualización simultánea requiere una interrupción temporal del servicio inferior a dos minutos durante el proceso de actualización mientras se reinician los nodos.

Actualización gradual

Una actualización gradual actualiza y reinicia cada nodo de forma individual en el clúster, uno tras otro. Durante una actualización gradual, el clúster permanece en línea y continúa suministrando datos a los clientes sin que el servicio se interrumpa. Antes de OneFS 8.0, una actualización gradual solo podía realizarse dentro de una familia de versiones de código de OneFS y no entre las revisiones de versión de código principal de OneFS. De OneFS 8.0 en adelante, todas las versiones nuevas se actualizarán a partir de la versión anterior.

Actualizaciones no disruptivas

Las actualizaciones no disruptivas (NDU) permiten que un administrador de clústeres actualice el SO de almacenamiento mientras los usuarios finales continúan accediendo a los datos sin errores ni interrupciones. La actualización del sistema operativo en un clúster es sencillamente una cuestión de actualización gradual. Durante este proceso, se actualiza un nodo a la vez al código nuevo, y los clientes NFS y SMB3 activos conectados a él se migran automáticamente a otros nodos del clúster. También se permite una actualización parcial, en la cual se puede actualizar un subconjunto de nodos del clúster. El subconjunto de nodos también puede crecer durante la actualización. Una actualización permite pausar y reanudar una actualización, con lo cual los clientes pueden distribuir las actualizaciones en varias ventanas de mantenimiento más pequeñas. Además, OneFS 8.2.2 y posteriores ofrecen actualizaciones paralelas, mediante las cuales los clústeres pueden actualizar un vecindario completo, o dominio de fallas, a la vez, lo que reduce considerablemente la duración de las actualizaciones de clústeres grandes. OneFS 9.2 y versiones posteriores combinan las actualizaciones de sistema operativo y firmware, lo que reduce considerablemente el impacto y la duración de las actualizaciones, ya que permiten que se realicen en conjunto. 9.2 y versiones posteriores también incluyen actualizaciones basadas en vaciado, en las que se impide que los nodos se reinicien o reinicien los servicios de protocolo hasta que todos los clientes SMB se desconecten del nodo.

Capacidad de reversión

OneFS admite la reversión de actualizaciones, lo que proporciona la capacidad de devolver un clúster con una actualización sin confirmar a su versión anterior de OneFS.

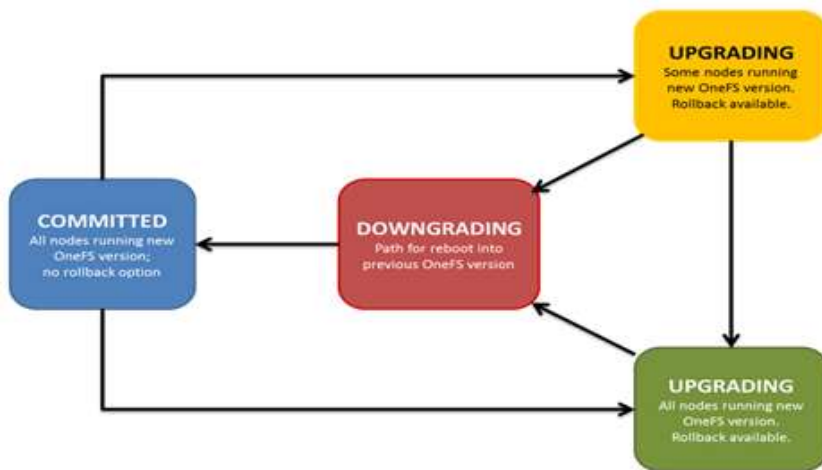


Figura 28: estados de actualización no disruptiva de OneFS

Actualizaciones automáticas de firmware

Los clústeres con tecnología OneFS admiten las actualizaciones automáticas de firmware de unidades nuevas y de reemplazo como parte del proceso de actualización de firmware no disruptivo. Las actualizaciones de firmware se proporcionan a través de paquetes de compatibilidad de unidades, que simplifican y optimizan la administración de las unidades existentes y nuevas en todo el clúster. Esto garantiza que el firmware de la unidad esté actualizado y reduce la probabilidad de fallas debido a problemas conocidos de la unidad. Por lo tanto, las actualizaciones automáticas de firmware de unidad son un componente importante de la estrategia de operaciones no disruptivas y de alta disponibilidad de OneFS. El firmware de unidad y de nodo se puede aplicar como una actualización gradual o mediante un reinicio completo del clúster.

Antes de OneFS 8.2, las actualizaciones de firmware de los nodos tenían que instalarse un nodo a la vez mediante una operación que consumía mucho tiempo, especialmente en clústeres de gran tamaño. Las actualizaciones de firmware de nodo ahora se pueden coreografiar en todo el clúster mediante la entrega de una lista de nodos que se actualizarán simultáneamente. La herramienta de ayuda de actualización se puede usar para seleccionar una combinación de nodos deseada que se puede actualizar simultáneamente y una lista explícita de nodos que no se deben actualizar juntos (por ejemplo, los nodos en un par de nodos).

Ejecución de la actualización

Como parte de una actualización, OneFS ejecuta automáticamente un control de verificación previo a la instalación. Esto verifica que la configuración de la instalación actual de OneFS sea compatible con la versión de OneFS que está destinada a la actualización. Cuando se encuentra una configuración no compatible, se detiene la actualización y se muestran instrucciones para solucionar el problema. Ejecutar proactivamente la comprobación de la actualización previa a la instalación antes de iniciar una actualización ayuda a evitar cualquier interrupción debido a una configuración incompatible.

Software de administración y protección de datos de OneFS

OneFS ofrece un portafolio integral de software de administración y protección de datos para satisfacer sus necesidades:

Módulo de software	Función	Descripción
CloudIQ™	Monitoreo del estado del clúster	Implementar el análisis inteligente y predictivo para monitorear proactivamente el estado del clúster.
InsightIQ™	Administración del rendimiento	Maximice el rendimiento de clúster con innovadoras herramientas de monitoreo e informes de rendimiento
DataIQ™	Administración y análisis de datos	Ubique datos, acceda a ellos y adminístrelos en segundos sin importar dónde residan, ya sea en el almacenamiento de archivos, en las instalaciones o en la nube. Obtenga una vista integral de los sistemas de almacenamiento heterogéneos con un solo panel y elimine de manera eficaz los datos capturados en silos.
SmartPools™	Administración de recursos	Implemente una estrategia altamente eficiente de almacenamiento en niveles automatizado para optimizar los costos y el rendimiento de almacenamiento
SmartQuotas™	Administración de datos	Asigne y administre cuotas que, de forma transparente, particionen el almacenamiento y lo aprovisionen de manera delgada en segmentos fácilmente administrados en los niveles de clúster, directorio, subdirectorio, usuario y grupo
SmartConnect™	Acceso a datos	Habilite el balanceo de carga de la conexión de los clientes, además de la conmutación por error y la conmutación por recuperación dinámicas de NFS para las conexiones de los clientes en todos los nodos de almacenamiento a fin de optimizar el uso de los recursos del clúster
SnapshotIQ™	Protección de datos	Proteja los datos de manera eficiente y confiable con instantáneas seguras casi inmediatas, con una sobrecarga del rendimiento mínima o nula. Acelere la recuperación de datos cruciales con restauraciones de instantáneas según demanda casi inmediatas. Cree copias modificables y con uso eficiente del espacio y el tiempo de una instantánea de solo lectura con instantáneas con capacidad de escritura de OneFS.
SynclQ™	Replicación de datos	Replice y distribuya grandes conjuntos de datos de misión crítica de manera asíncrona para multiplicar los sistemas de almacenamiento compartido en varios sitios a fin de lograr una funcionalidad recuperación ante desastres confiable. Sencillez de conmutación por error y conmutación por recuperación fácil de usar para aumentar la disponibilidad de los datos de misión crítica.
SmartLock™	Retención de datos	Proteja los datos cruciales contra la modificación o la eliminación accidental, prematura o maliciosa con nuestro enfoque de Write Once, Read Many (WORM) basado en software y satisfaga las estrictas necesidades de cumplimiento de normas y buen manejo y control tales como los requisitos de SEC 17a-4.
SmartDedupe™	Desduplicación de datos	Maximice la eficiencia del almacenamiento mediante el escaneo del clúster en busca de bloques idénticos y la eliminación de los duplicados, lo que reduce la cantidad de almacenamiento físico necesario.
CloudPools™	Organización de la nube en niveles	CloudPools le permite definir qué datos de su clúster deben archivar en el almacenamiento de nube. Los proveedores de nube incluyen Microsoft Azure, Google Cloud, Amazon S3, Dell EMC ECS y OneFS nativa.

Tabla 3: portafolio de servicios de datos de escala de alimentación de Dell EMC

Consulte la documentación del producto para obtener más información.

Conclusión

Con las soluciones NAS de escalamiento horizontal de Dell EMC con tecnología del sistema operativo OneFS, las organizaciones pueden escalar de TB a PB dentro de un solo sistema de archivos, un solo volumen, con un punto único de administración. OneFS ofrece alto rendimiento, alta producción o ambos sin tener que agregar más complejidad de administración.

Los centros de datos de última generación deben estar creados para ofrecer una escalabilidad sostenible. Utilizarán el poder de la automatización; aprovecharán el consumo masivo de hardware, garantizarán el consumo completo del fabric de red y proporcionarán máxima flexibilidad para respaldar el intento de las organizaciones por satisfacer un conjunto de requisitos en constante cambio.

OneFS es el sistema de archivos de última generación diseñado para enfrentar estos retos. OneFS proporciona lo siguiente:

- Un solo sistema de archivos completamente distribuido
- Clúster de alto rendimiento completamente simétrico
- Fraccionado de archivos en todos los nodos de un clúster
- Software automatizado para eliminar la complejidad
- Balanceo dinámico de contenido
- Protección de datos flexible
- Alta disponibilidad
- Administración basada en la web y en la línea de comandos

OneFS es ideal para aplicaciones de “big data” basadas en archivos y no estructuradas en entornos empresariales de lago de datos, incluidos directorios principales, recursos compartidos de archivos, archivos, virtualización y análisis del negocio a gran escala, así como para una amplia variedad de entornos de computación de alto rendimiento con gran uso de datos, incluida la exploración de energía, los servicios financieros, los servicios de Internet y hosting, la inteligencia comercial, la ingeniería, la fabricación, los medios de comunicación y entretenimiento, la bioinformática y la investigación científica.

DÉ UN PASO ADELANTE

Póngase en contacto con su representante de ventas de Dell EMC o con su reseller autorizado para obtener más información acerca de cómo las soluciones de almacenamiento NAS de PowerScale pueden beneficiar a su organización.

[Visite Dell EMC PowerScale](#) para comparar las características y obtener más información.



Más información sobre las soluciones Dell EMC PowerScale



Comunicarse con un experto de Dell EMC



Ver más recursos



Únase a la conversación con #DellEMCStorage