DELL Technologies

WHITE PAPER

DELL EMC POWERSCALE ONEFS: UMA VISÃO GERAL TÉCNICA

Resumo geral

Este white paper apresenta detalhes técnicos sobre os principais recursos do sistema operacional OneFS que são usados para alimentar todas as soluções de armazenamento NAS de scale-out do Dell EMC PowerScale.

Setembro de 2021

Revisões

Versão	Data	Comentário
1.0	Novembro de 2013	Versão inicial do OneFS 7.1
2.0	Junho de 2014	Atualizado para o OneFS 7.1.1
3.0	Novembro de 2014	Atualizado para o OneFS 7.2
4.0	Junho de 2015	Atualizado para o OneFS 7.2.1
5.0	Novembro de 2015	Atualizado para o OneFS 8.0
6.0	Setembro de 2016	Atualizado para o OneFS 8.0.1
7.0	Abril de 2017	Atualizado para o OneFS 8.1
8.0	Novembro de 2017	Atualizado para o OneFS 8.1.1
9.0	Fevereiro de 2019	Atualizado para o OneFS 8.1.3
10.0	Abril de 2019	Atualizado para o OneFS 8.2
11.0	Agosto de 2019	Atualizado para o OneFS 8.2.1
12.0	Dezembro de 2019	Atualizado para o OneFS 8.2.2
13.0	Junho de 2020	Atualizado para o OneFS 9.0
14.0	Setembro de 2020	Atualizado para o OneFS 9.1
15.0	Abrilde 2021	Atualizado para o OneFS 9.2
16.0	Setembro de 2021	Atualizado para o OneFS 9.3

Agradecimentos

Este artigo foi produzido por:

Autor: Nick Trimbee

ÍNDICE

Introdução 4
Visão geral do OneFS 4
Nós do PowerScale 5
Rede 6
Visão geral do software OneFS 7
Estrutura do file system 10
Layout de dados 11
Gravações de arquivos 12
Armazenamento em cache do OneFS 15
Coerência de cache do OneFS 17
Cache de nível 1 17
Cache de nível 2 18
Cache de nível 3 18
Leituras de arquivo 20
Bloqueios e simultaneidade 21
E/S com múltiplos threads 22
Proteção de dados 23
Compatibilidade 30
Protocolos compatíveis 31
Operações não disruptivas – suporte a protocolos 32
Filtragem de arquivos 32
Desduplicação de dados – SmartDedupe 32
Eficiência de armazenamento de pequenos arquivos 34
Redução de dados em linha 34
Interfaces 37
Autenticação e controle de acesso 37
Active Directory 38
Zonas de acesso 38
Administração com base em funções 39
Auditoria do OneFS 39
Upgrade de software 39
Software OneFS de gerenciamento e proteção de dados 41
Conclusão 42
DÊ O PRÓXIMO PASSO 42

Introdução

As três camadas do modelo de armazenamento tradicional – file system, gerenciador de volumes e proteção de dados – evoluíram ao longo do tempo para atender às necessidades de arquiteturas de armazenamento em pequena escala, mas apresentam complexidade significativa e não são bem adaptadas para sistemas em escala de petabyte. O sistema operacional OneFS substitui todos esses, fornecendo um sistema unificador de arquivos em cluster com proteção de dados escalável integrada e eliminando a necessidade de gerenciamento de volume. O OneFS é um componente básico fundamental de infraestruturas de scale-out, permitindo dimensionamento em grande escala e tremenda eficiência, e é usado para alimentar todas as soluções de armazenamento NAS de scale-out do Dell EMC PowerScale.

Fundamentalmente, o OneFS foi projetado para ser dimensionado não só em termos de máquinas, mas também em termos humanos, permitindo que sistemas de grande escala sejam gerenciados com uma fração do pessoal necessário para os sistemas tradicionais de armazenamento. O OneFS elimina a complexidade e incorpora a funcionalidade de autocorreção e autogerenciamento que reduz drasticamente a carga de gerenciamento de armazenamento. O OneFS também incorpora paralelismo em um profundo nível do sistema operacional de tal forma que praticamente todos os serviços de sistema importantes sejam distribuídos em várias unidades de hardware. Isso permite que o OneFS seja dimensionado em praticamente todas as dimensões conforme a infraestrutura é expandida, garantindo que o que funciona hoje continue funcionando à medida que o conjunto de dados aumenta.

O OneFS é um file system totalmente simétrico com nenhum ponto único de falha – aproveitando a organização por clusters não só para aumentar o desempenho e a capacidade, mas também para permitir failover de qualquer ponto a qualquer ponto e múltiplos níveis de redundância que vão muito além dos recursos de RAID. A tendência para subsistemas de discos tem sido o desempenho com aumento lento enquanto as densidades de armazenamento aumentam rapidamente. O OneFS responde a essa realidade com a expansão da quantidade de redundância, bem como a velocidade de reparação de falhas. Isso permite que o OneFS aumente para uma escala de vários petabytes, proporcionando mais confiabilidade do que pequenos sistemas tradicionais de armazenamento.

O hardware do PowerScale fornece o equipamento no qual o OneFS é executado. Os componentes de hardware são avançados, mas baseados em hardware genérico – garantindo os benefícios do custo e das curvas de eficiência cada vez melhores desse tipo de hardware. O OneFS permite que um hardware seja incorporado ou removido do cluster à vontade e a qualquer momento, abstraindo os dados e os aplicativos do hardware. Os dados recebem longevidade infinita, são protegidos contra as vicissitudes das gerações de hardware em evolução. O custo, bem como os problemas de migrações de dados e atualizações de hardware são eliminados.

O OneFS é idealmente adequado para aplicativos de Big Data não estruturado baseados em arquivos em ambientes corporativos, inclusive diretórios base de grande escala, compartilhamentos de arquivos, arquivamentos, virtualização e lógica analítica de negócios. Como tal, o OneFS é amplamente utilizado em muitos setores com uso intenso de dados hoje em dia, incluindo energia, serviços financeiros, de Internet e serviços de hospedagem, business intelligence, engenharia, manufatura, mídia e entretenimento, bioinformática, investigação científica e outros ambientes computacionais de alto desempenho.

Público-alvo

Este artigo apresenta informações sobre a implementação e o gerenciamento de clusters do Dell EMC PowerScale e oferece um histórico abrangente para a arquitetura do OneFS.

O público-alvo deste white paper é qualquer pessoa que esteja configurando e gerenciando um ambiente de armazenamento em cluster do PowerScale. Presume-se que o leitor tenha noções básicas de armazenamento, sistema de rede, sistemas operacionais e gerenciamento de dados.

📖 Mais informações sobre a configuração de comandos e recursos do OneFS estão disponíveis no Guia de administração do OneFS.

Visão geral do OneFS

O OneFS combina as três camadas de arquiteturas tradicionais de armazenamento — file system, gerenciador de volumes e proteção de dados — em uma camada de software unificado, criando um file system distribuído único e inteligente que é executado em um cluster de armazenamento do OneFS.

Figura 1: o OneFS combina file system, gerenciador de volumes e proteção de dados em um só sistema distribuído e inteligente.

Esta é a principal inovação que permite diretamente que as empresas utilizem com sucesso o NAS de scale-out em seus ambientes de hoje. Ele adere aos princípios fundamentais de scale-out, software inteligente, hardware genérico e arquitetura distribuída. O OneFS não é apenas o sistema operacional, mas também o file system subjacente que impulsiona e armazena dados no cluster.

Nós do PowerScale

O OneFS trabalha exclusivamente com os nós de plataforma dedicados chamados de “cluster”. Um cluster único consiste em vários nós, que são construídos como equipamentos corporativos montáveis em rack que contêm contendo memória, CPU, sistema de rede, Ethernet ou interconexões InfiniBand de baixa latência, controladoras de disco e mídia de armazenamento. Como tal, cada nó no cluster distribuído tem recursos de computação, bem como recursos de armazenamento ou capacidade.

Com a arquitetura de 6ª geração, um único chassi de 4 nós em um formato de 4RU (unidades de rack) é necessário para criar um cluster que pode ser dimensionado para até 252 nós no OneFS 8.2 e posterior. As plataformas de nós individuais precisam de, no mínimo, 3 nós e 3RU de espaço em rack para formar um cluster. Há diversos tipos diferentes de nós, todos os quais podem ser incorporados em um só cluster no qual diferentes nós proporcionam taxas variadas de capacidade para throughput ou IOPS. Os nós independentes do chassi de 6ª geração tradicional e do PowerScale All-Flash F900, F600 e F200 vão coexistir no mesmo cluster.

Cada nó ou chassi adicionado ao cluster aumenta a capacidade agregada de disco, cache, CPU e rede. O OneFS utiliza cada um dos componentes básicos de hardware a fim de que o todo se torne maior do que a soma das partes. A memória RAM é agrupada em um só cache coerente, permitindo E/S em qualquer parte do cluster para beneficiar-se de dados em cache em qualquer lugar. Um registro do file system garante que as gravações sejam seguras em caso de falta de energia. Os spindles e a CPU são combinados para aumentar o throughput, a capacidade e a IOPS, à medida que o cluster aumenta, para fins de acesso a um arquivo ou vários arquivos. A capacidade de armazenamento de um cluster pode variar de dezenas de TBs a dezenas de PBs. A capacidade máxima vai continuar aumentando à medida que a mídia de armazenamento e o chassi do nó continuam ficando mais densos.

Os nós da plataforma com tecnologia do OneFS são divididos em várias classes ou níveis, de acordo com sua funcionalidade:

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

Tabela 1: níveis de hardware e tipos de nós

Rede

Existem dois tipos de redes associados a um cluster: interna e externa.

Rede de back-end

Todas as comunicações entre nós em um cluster são realizadas em uma rede de back-end dedicada, composta por Ethernet de 10, 40 ou 100 Gb ou InfiniBand (IB) QDR de baixa latência. Esta rede de back-end, que é configurada com switches redundantes para proporcionar alta disponibilidade, atua como o backplane para o cluster. Isso permite que cada nó atue como colaborador no cluster e isole a comunicação de nó para nó em uma rede privada, de alta velocidade e baixa latência. Essa rede de back-end utiliza IP (Internet Protocol) para a comunicação entre nós.

Rede front-end

Os clients se conectam ao cluster usando conexões Ethernet (10 GbE, 25 GbE, 40 GbE ou 100 GbE) que estão disponíveis em todos os nós. Como cada nó fornece suas próprias portas Ethernet, a quantidade de largura de banda de rede disponível para o cluster é dimensionada de forma linear com desempenho e capacidade. Um cluster aceita protocolos de comunicação de rede padrão para uma rede do cliente, incluindo NFS, SMB, HTTP, FTP, HDFS e S3. Além disso, o OneFS oferece integração total com ambientes IPv4 e IPv6.

Visualização completa do cluster

O cluster completo é combinado com hardware, software, redes na seguinte exibição:

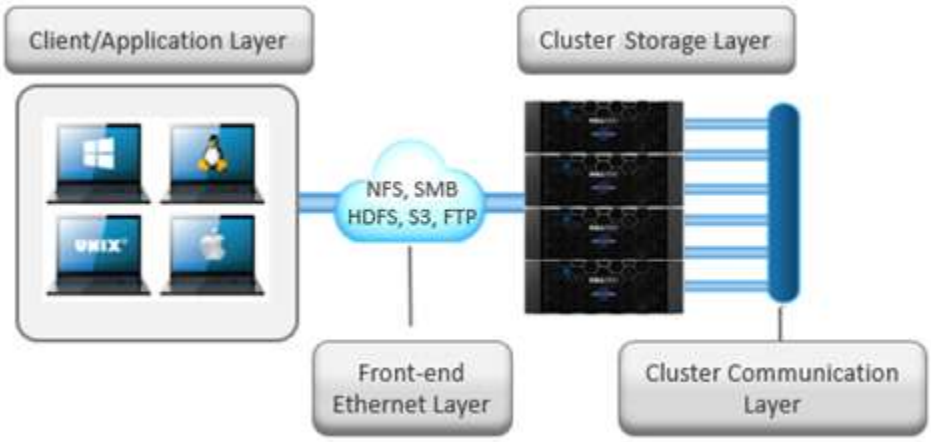

Figura 2: todos os componentes do OneFS em funcionamento

O diagrama acima mostra a arquitetura completa, hardware, software e rede, todos trabalhando juntos em seu ambiente com servidores para fornecer um só file system completamente distribuído que pode se dimensionar de forma dinâmica à medida que as necessidades de cargas de trabalho e capacidade ou de throughput mudam em um ambiente de scale-out.

O OneFS SmartConnect é um balanceador de carga que funciona na camada Ethernet de front-end para distribuir uniformemente as conexões do client em todo o cluster. O SmartConnect dá suporte a failover e failback dinâmicos do NFS para clients Linux e UNIX, bem como disponibilidade contínua do SMB3 para clients Windows. Isso garante que, quando ocorre falha em um nó, ou quando é realizada uma manutenção preventiva, todas as gravações e leituras em trânsito são passadas para outro nó no cluster para concluir sua operação sem qualquer interrupção para o aplicativo ou usuário.

Durante um failover, os clients são redistribuídos igualmente entre todos os nós restantes no cluster para garantir o menor impacto possível sobre o desempenho. Se um nó é desativado por qualquer motivo, incluindo uma falha, os endereços IP virtuais nesse nó são perfeitamente migrados para outro nó no cluster. Quando o nó off-line é colocado novamente on-line, o SmartConnect reequilibra automaticamente os clients NFS e SMB3 por todo o cluster para garantir máxima utilização de desempenho e armazenamento. Para manutenções periódicas do sistema e atualizações de software, essa funcionalidade permite rolling upgrades por nó que oferecem alta disponibilidade por toda a duração da janela de manutenção.

📖 Mais informações estão disponíveis no white paper OneFS SmartConnect.

Visão geral do software OneFS

Sistema operacional

O OneFS é desenvolvido sobre um sistema operacional UNIX baseado em BSD. Ele aceita nativamente tanto semântica Linux/Unix quanto Windows, incluindo vínculos físicos, delete-on-close, renomeação atômica, ACLs e atributos estendidos. Ele usa o BSD como base de seu sistema operacional, pois é um sistema operacional maduro e comprovado, e a comunidade de código aberto pode ser aproveitada para fins de inovação. Do OneFS 8.2 em diante, a versão subjacente do sistema operacional é o FreeBSD 11.

Serviços de client

Os protocolos de front-end que os clients podem usar para interagir com o OneFS são chamados de serviços de client. Consulte a seção Supported Protocols para obter uma lista detalhada de protocolos compatíveis. Para entender, como o OneFS se comunica com os clients, dividimos o subsistema de E/S em duas metades: a metade superior ou o “iniciador” e a metade inferior ou o “participante”. Cada nó do cluster é um participante de uma operação de E/S em particular. O nó a que o client se conecta é o iniciador e esse nó atua como o “capitão” de toda a operação de E/S. As operações de leitura e gravação são detalhadas em seções posteriores

Operações de cluster

Em uma arquitetura de cluster, há trabalhos de cluster, que são responsáveis por cuidar da integridade e da manutenção do próprio cluster – esses trabalhos são todos gerenciados através do OneFS. O Job Engine é executado em todo o cluster e é responsável pela divisão e a conquista de grandes tarefas de proteção e gerenciamento de armazenamento. Para conseguir isso, ele reduz uma tarefa em itens de trabalho menores e, em seguida, aloca, ou mapeia, essas partes do trabalho geral para vários threads de operador em cada nó. O progresso é rastreado e relatado durante a execução do trabalho, além disso, um relatório detalhado e o status são apresentados após a conclusão ou o encerramento.

O Job Engine inclui um sistema abrangente e de checkpointing para verificação, que permite que os trabalhos sejam pausados e retomados, além de interrompidos e iniciados. O framework do Job Engine também inclui um sistema adaptável de gerenciamento de impacto.

Normalmente, o Job Engine executa trabalhos como tarefas de segundo plano em todo o cluster, usando recursos e capacidade sobressalentes ou especialmente reservados. Os trabalhos em si podem ser categorizados em três classes primárias:

Trabalhos de manutenção do file system

Esses trabalhos realizam a manutenção do file system em segundo plano e, normalmente, exigem acesso a todos os nós. Esses trabalhos são necessários para serem executados em configurações padrão e, muitas vezes, em condições degradadas do cluster. Os exemplos incluem proteção de file system e reconstruções de unidades.

Trabalhos de suporte a recursos

Os trabalhos de suporte a recursos realizam trabalho que facilita algumas funções de gerenciamento de armazenamento estendido e, normalmente, são executados apenas quando o recurso foi configurado. Entre os exemplos estão a desduplicação e a varredura antivírus.

Trabalhos de ação do usuário

Esses trabalhos são executados diretamente pelo administrador de armazenamento para atingir alguma meta de gerenciamento de dados. Os exemplos incluem exclusões de árvores paralelas e manutenção de permissões.

A tabela a seguir apresenta uma lista abrangente dos trabalhos expostos do Job Engine, as operações que eles executam e seus respectivos métodos de acesso ao file system:

Nome do trabalho	Descrição do trabalho	Método de acesso
AutoBalance	Equilibra o espaço livre no cluster.	**Unidade + LIN**
AutoBalanceLin	Equilibra o espaço livre no cluster.	**LIN**
AVScan	Trabalho de varredura de vírus que os servidores antivírus executam.	**Árvore**
ChangelistCreate	Criar uma lista de alterações entre dois snapshots consecutivos do SyncIQ	**Changelist**
CloudPoolsLin	Arquiva os dados em um provedor de serviços em nuvem de acordo com uma política de pool de arquivos.	**LIN**
CloudPoolsTreewalk	Arquiva os dados em um provedor de serviços em nuvem de acordo com uma política de pool de arquivos.	**Árvore**
Coletar	Recupera o espaço em disco que não pode ser liberado devido a um nó ou uma unidade estar disponível enquanto eles sofrem de várias condições de falha.	**Unidade + LIN**
ComplianceStoreDelete	Trabalho de coleta de lixo do modo de conformidade SmartLock.	**Árvore**
Desduplicação	Desduplica blocks idênticos no file system.	**Árvore**
DedupeAssessment	Realiza a avaliação dry-run dos benefícios da desduplicação.	**Árvore**
DomainMark	Associa um caminho e seu conteúdo a um domínio.	**Árvore**
DomainTag	Associa um caminho e seu conteúdo a um domínio.	**Árvore**
EsrsMftDownload	Trabalho de transferência de arquivos gerenciados do ESRS para arquivos de licença.	
FilePolicy	Trabalho eficiente de política de pools de arquivos do SmartPools.	**Changelist**
FlexProtect	Reconstrói e protege novamente o file system para se recuperar de um cenário de falha.	**Unidade + LIN**
FlexProtectLin	Protege novamente o file system.	**LIN**
FSAnalyze	Reúne dados da análise lógica do file system que são usados em conjunto com o InsightIQ.	**Changelist**
IndexUpdate	Cria e atualiza um índice eficiente de file system para os trabalhos FilePolicy e FSAnalyze.	**Changelist**
IntegrityScan	Realiza a correção e a verificação on-line de qualquer inconsistência no file system.	**LIN**
LinCount	Varre e conta os LINs (Logical Inodes, inodes lógicos) do file system.	**LIN**
MediaScan	Analisa unidades em busca de erros no nível de mídia.	**Unidade + LIN**
MultiScan	Executa trabalhos Collect e AutoBalance simultaneamente.	**LIN**

Nome do trabalho	Descrição do trabalho	Método de acesso
PermissionRepair	Corrige permissões de arquivos e diretórios.	**Árvore**
QuotaScan	Atualiza a contabilidade de quota para domínios criados em um caminho de diretório existente.	**Árvore**
SetProtectPlus	Aplica a política de arquivo padrão. Esse trabalho é desativado se o SmartPools está ativado no cluster.	**LIN**
ShadowStoreDelete	Libera espaço associado a um armazenamento de sombra.	**LIN**
ShadowStoreProtect	Protege os armazenamentos de sombra que são referidos por um LIN com mais proteção solicitada.	**LIN**
ShadowStoreRepair	Repara os armazenamentos de sombra.	**LIN**
SmartPools	Trabalho que executa e move dados entre os níveis de nós no mesmo cluster. Também executa a funcionalidade do CloudPools se estiver licenciada e configurada.	**LIN**
SmartPoolsTree	Impõe as políticas de arquivos do SmartPools em uma subárvore.	**Árvore**
SnapRevert	Inverte todo um snapshot.	**LIN**
SnapshotDelete	Libera espaço em disco que está associado a snapshots excluídos.	**LIN**
TreeDelete	Exclui um caminho no file system diretamente do próprio cluster.	**Árvore**
Undedupe	Remove a desduplicação de blocks idênticos no file system.	**Árvore**
Upgrade	Faz upgrade do cluster em uma versão mais recente do OneFS.	**Árvore**
WormQueue	Analisa a fila de LIN do SmartLock	**LIN**

Figura 1: descrições de trabalho do OneFS Job Engine

Embora os trabalhos de manutenção do file system sejam executados por padrão, seja em um agendamento ou em reação a um evento específico do file system, qualquer trabalho do Job Engine pode ser gerenciado por meio da configuração de seu nível de prioridade (em relação a outros trabalhos) e de sua política de impacto.

Uma política de impacto pode consistir em um ou muitos intervalos de impacto, que são blocos de tempo dentro de uma determinada semana. Cada intervalo de impacto pode ser configurado para usar um só nível de impacto predefinido que especifica que quantidade de recursos de cluster usar para uma operação de cluster em particular. Os níveis de impacto disponíveis do Job Engine são:

- Paused
- Baixo
- Médio
- Alto

Esse grau de granularidade permite que os intervalos e os níveis de impacto sejam configurados por trabalho, a fim de garantir uma operação tranquila do cluster. As políticas de impacto resultantes determinam quando um trabalho é executado e os recursos que um trabalho pode consumir.

Além disso, os trabalhos do Job Engine são priorizados em uma escala de um a dez, com um valor menor representando uma prioridade mais alta. Isso é semelhante ao conceito do utilitário de agendamento do UNIX, “bom”.

O Job Engine permite que até três trabalhos sejam executados simultaneamente. Essa execução de trabalho simultânea é regida pelos seguintes critérios:

- Prioridade do trabalho
- Conjuntos de exclusão — trabalhos que não podem ser executados juntos (ou seja, FlexProtect e AutoBalance)
- Integridade do cluster – a maioria dos trabalhos não pode ser executada quando o cluster está em um estado degradado.

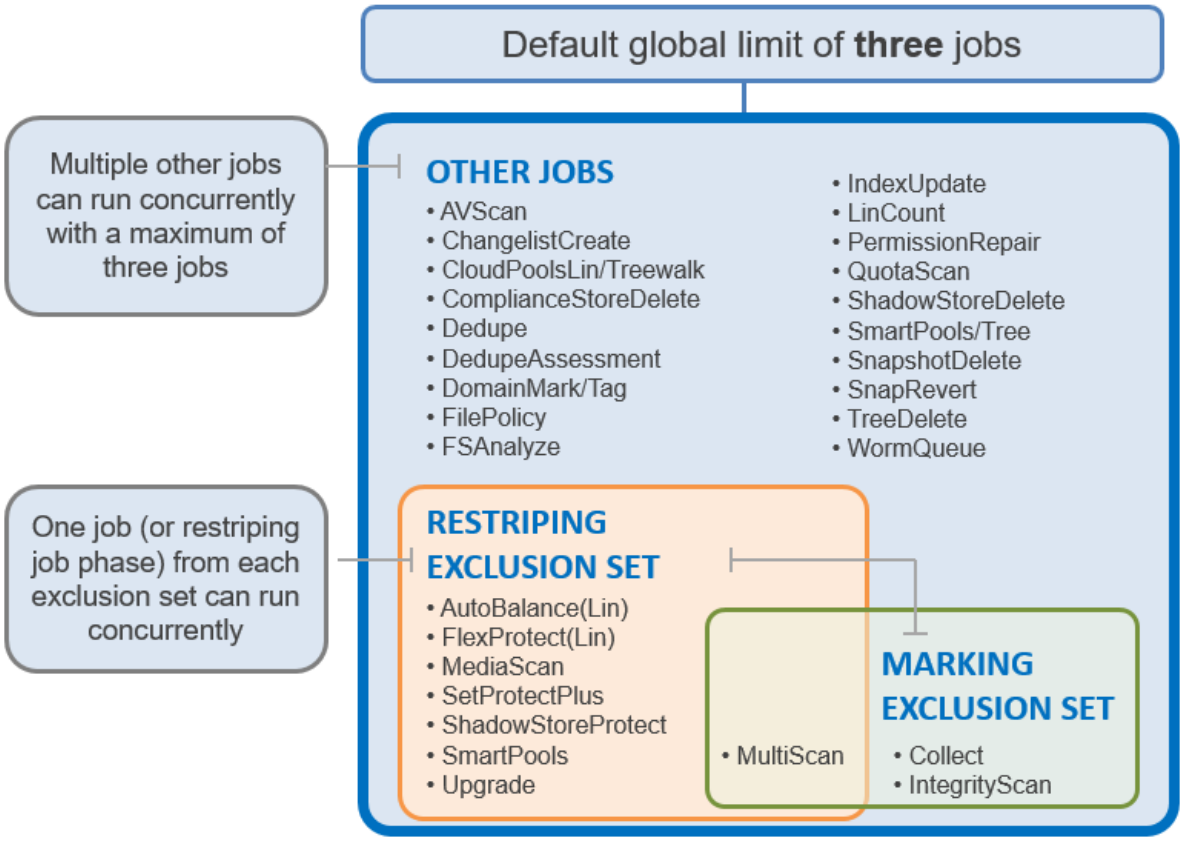

Figura 4: conjuntos de exclusão do OneFS Job Engine

📖 Mais informações estão disponíveis no white paper OneFS Job Engine.

Estrutura do file system

O file system do OneFS baseia-se no UFS (Unix file system, file system do UNIX) e, portanto, é um file system distribuído muito rápido. Cada cluster cria um namespace e um file system. Isso significa que o file system é distribuído em todos os nós do cluster e é acessível por clients que se conectam a qualquer nó no cluster. Não há particionamento nem a necessidade de criação de volume. Em vez de limitar o acesso ao espaço livre e a arquivos não autorizados no nível de volume físico, o OneFS prevê a mesma funcionalidade no software através de permissões de compartilhamentos e arquivos, e através do serviço SmartQuotas, que fornece gerenciamento de quotas em nível de diretório.

🕮 Mais informações estão disponíveis no white paper OneFS SmartQuotas.

Como todas as informações são compartilhadas entre nós em toda a rede interna, os dados podem ser gravados ou lidos a partir de qualquer nó, otimizando o desempenho quando vários usuários estão lendo e gravando simultaneamente no mesmo conjunto de dados.

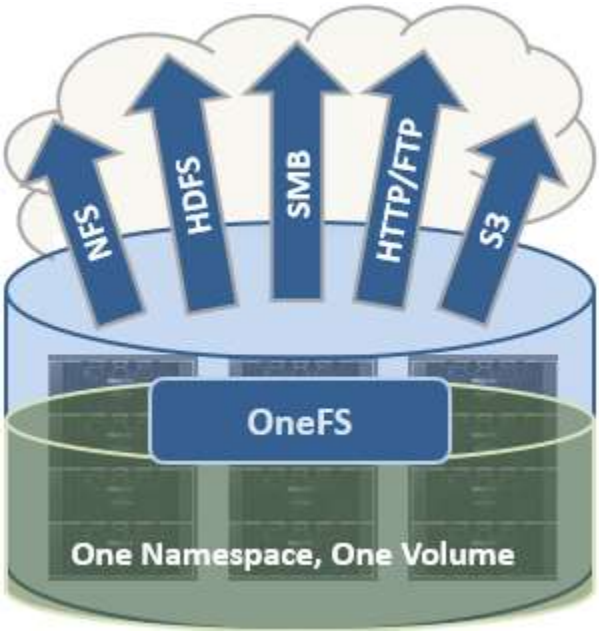

Figura 5: file system único com vários protocolos de acesso

O OneFS é verdadeiramente um file system único com um namespace. Os dados e metadados são distribuídos entre os nós para fins de redundância e disponibilidade. O armazenamento foi completamente virtualizado para os usuários e o administrador. A árvore de arquivos pode crescer organicamente sem a necessidade de planejamento ou supervisão sobre como a árvore cresce ou como os usuários a usam. Nenhum estudo especial tem de ser aplicado pelo administrador sobre classificação de arquivos por níveis para o disco apropriado porque o OneFS SmartPools vai lidar com isso automaticamente, sem interromper a única árvore. Nenhuma consideração especial precisa ser dada à forma como se pode replicar tal árvore grande porque o serviço OneFS SyncIQ paraleliza automaticamente a transferência da árvore de arquivos para um ou mais clusters alternativos, sem ter em conta a forma nem a profundidade da árvore de arquivos.

Esse projeto deve ser comparado à agregação de namespaces, que é uma tecnologia comumente usada para fazer o NAS tradicional “parecer” que tem um só namespace. Com a agregação de namespaces, os arquivos ainda têm de ser gerenciados em volumes separados, mas uma simples camada de “verniz” permite que diretórios individuais em volumes sejam “colados” a uma árvore de “nível superior” através de links simbólicos. Nesse modelo, LUNs e volumes, bem como os limites de volumes, ainda estão presentes. Os arquivos têm de ser movidos manualmente de volume a volume a fim de realizar o balanceamento de carga. O administrador tem de ser cuidadoso sobre como a árvore é exibida. O armazenamento em camadas está longe de ser perfeito e requer intervenção significativa e contínua. O failover requer espelhamento de arquivos entre volumes, reduzindo a eficiência e aumentando os custos de compra, energia e refrigeração. Os encargos gerais do administrador ao usar agregação de namespace são maiores do que para um dispositivo NAS simples e tradicional. Isso evita que tais infraestruturas cresçam demais.

Layout de dados

O OneFS utiliza extensões e indicadores físicos para metadados e armazena metadados de arquivos e diretórios em inodes. Os LINs (Logical Inodes, inodes lógicos) do OneFS geralmente têm 512 bytes de tamanho, o que permite que eles se encaixem nos setores nativos nos quais a maioria das unidades de disco rígido é formatada. O suporte também é oferecido para inodes de 8 KB a fim de dar suporte às classes mais densas de discos rígidos, que agora são formatados com setores de 4 KB.

As B-trees são amplamente utilizadas no file system, permitindo dimensionamento para bilhões de objetos e pesquisas quase instantâneas de dados ou metadados. O OneFS é um file system completamente simétrico e altamente distribuído. Os dados e metadados são sempre redundantes em vários dispositivos de hardware. Os dados são protegidos usando a codificação de eliminação entre os nós do cluster; isso cria um cluster que tem alta eficiência, permitindo uma relação entre capacidade bruta e utilizável de 80% ou mais em grupos de cinco nós ou mais. Os metadados (que compõem geralmente menos de 1% do sistema) são espelhados no cluster para obter desempenho e disponibilidade. Como o OneFS não é dependente de RAID, a quantidade de redundância pode ser selecionada pelo administrador, no nível de arquivo ou diretório, para além dos padrões do cluster. O acesso a metadados e as tarefas de bloqueio são gerenciados por todos os nós, coletivamente e igualmente em uma arquitetura ponto a ponto. Essa simetria é a chave para a simplicidade e a capacidade de recuperação da arquitetura. Não há servidor de metadados, gerenciador de bloqueio ou nó gateway único.

Como o OneFS deve acessar blocks de diversos dispositivos simultaneamente, o esquema de endereçamento usado para dados e metadados é indexado no nível físico por um conjunto de valores de {nó, unidade, deslocamento}. Por exemplo, se 12345 fosse um endereço de block para um block que residisse no disco 2 do nó 3, a leitura seria {3.2.12345}. Todos os metadados dentro do cluster são triplamente espelhados (no mínimo) e sua proteção sempre está associada ao nível de proteção do arquivo. Por exemplo, se um arquivo estivesse com proteção de código contra eliminação de “+2n”, o que implica que o arquivo poderia aceitar duas falhas simultâneas, todos os metadados necessários para acessar esse arquivo seriam espelhados 3 vezes, por isso também poderia aceitar duas falhas. O file system inerentemente permite a qualquer estrutura usar todos e quaisquer blocks em todos os nós do cluster.

Outros sistemas de armazenamento enviam dados através de camadas de gerenciamento de volume e RAID, introduzindo ineficiências no layout de dados e fornecendo acesso não otimizado a blocks. O OneFS controla a colocação de arquivos diretamente, até o nível do setor em qualquer unidade em qualquer lugar do cluster. Isso permite padrões de colocação de dados otimizados e de E/S e evita operações desnecessárias de leitura-modificação-gravação. Colocando dados em discos arquivo por arquivo, o OneFS é capaz de controlar de modo flexível o tipo de particionamento, bem como o nível de redundância do sistema de armazenamento nos níveis de sistema, diretório e até mesmo de arquivo. Os sistemas tradicionais de armazenamento exigiriam que um volume de RAID inteiro fosse dedicado a um tipo especial de desempenho e configuração de proteção. Por exemplo, um conjunto de discos pode ser organizado em um nível de proteção de RAID 1+0 de uma base de dados. Isso torna difícil otimizar o uso do spindle sobre o conjunto inteiro de armazenamento (já que spindles ociosos não podem ser emprestados) e também leva a projetos inflexíveis que não se adaptam às necessidades dos negócios. O OneFS permite o ajuste individual e alterações flexíveis, a qualquer momento, totalmente on-line.

Gravações de arquivos

O software OneFS é executado em todos os nós igualmente – criando um só file system executado em cada nó. Não há um controle do nó ou “masters” no cluster; todos os nós são peers verdadeiros.

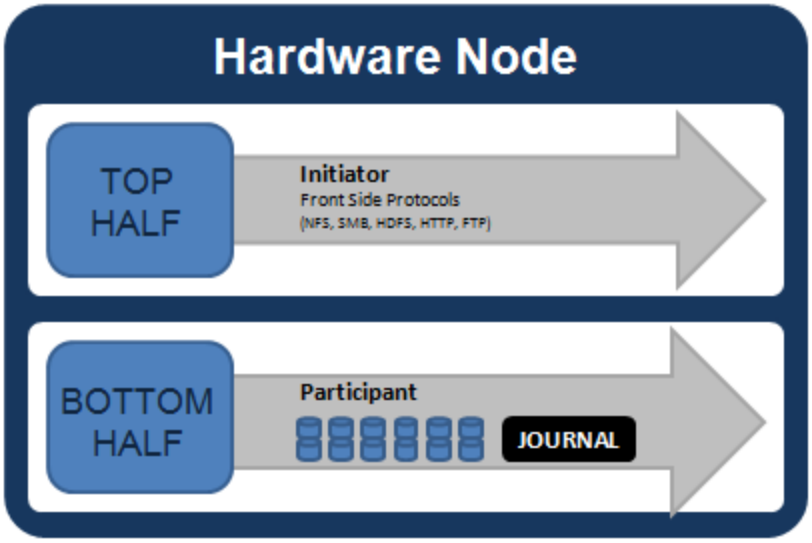

Figura 6: Modelo de componentes do nó envolvidos em E/S

Se fôssemos examinar todos os componentes dentro de cada nó de um cluster que estão envolvidos em E/S a partir de um alto nível, seria parecido com a Figura 6 acima. Dividimos o pacote de discos em uma camada “superior”, chamada de iniciador, e uma “inferior”, chamada de participante. Essa divisão é usada como “modelo lógico” para a análise de qualquer leitura ou gravação determinada. Em um nível físico, CPUs e cache de RAM nos nós estão simultaneamente lidando com tarefas de iniciador e participante para a E/S ocorrer em todo o cluster. Há caches e um gerenciador de bloqueio distribuído que são excluídos do diagrama acima para mantê-lo simples. Eles serão abordados em seções posteriores do documento.

Quando um client se conecta a um nó para gravar um arquivo, ele está se conectando à parte superior, ou iniciador, desse nó. Os arquivos são divididos em pequenos fragmentos lógicos chamados frações antes de serem gravados na metade inferior, ou participante, de um nó (disco). O buffer à prova de falha que utiliza um aglutinador de gravação é usado para assegurar que as gravações sejam eficientes e as operações de leitura-modificação-gravação sejam evitadas. O tamanho de cada fragmento de arquivo é conhecido como o tamanho da unidade de fração.

O OneFS fraciona os dados em todos os nós, e não simplesmente em discos, e protege os arquivos, diretórios e metadados associados através do código contra eliminação de software ou da tecnologia de espelhamento. Para dados, o OneFS pode usar (a critério do administrador) o sistema de codificação de eliminação Reed-Solomon para proteção de dados ou (menos comumente) o espelhamento. O espelhamento, quando aplicado aos dados do usuário, tende a ser mais utilizado para casos de transações de alto desempenho. A maior parte dos dados do usuário geralmente usa codificação de eliminação, pois proporciona um desempenho extremamente alto, sem sacrificar a eficiência em disco. A codificação de eliminação pode fornecer mais de 80% de eficiência no disco bruto com cinco nós ou mais, e em grandes clusters pode até fazê-lo ao mesmo tempo que proporciona redundância em nível quatro vezes maior. A largura da fração de um determinado arquivo é o número de nós (e não discos) em que um arquivo é gravado. Ela é determinada pelo número de nós no cluster, o tamanho do arquivo e a configuração de proteção (por exemplo, +2n).

O OneFS usa algoritmos avançados para determinar o layout dos dados a fim de obter eficiência e desempenho máximos. Quando um client se conecta a um nó, o iniciador desse nó age como o “capitão” do layout de dados de gravação desse arquivo. Dados, codificação de eliminação (ECC), metadados e inodes estão distribuídos em vários nós em um cluster e até mesmo em várias unidades dentro de nós.

A Figura 7 abaixo mostra uma gravação de arquivo acontecendo em todos os nós em um cluster de três nós.

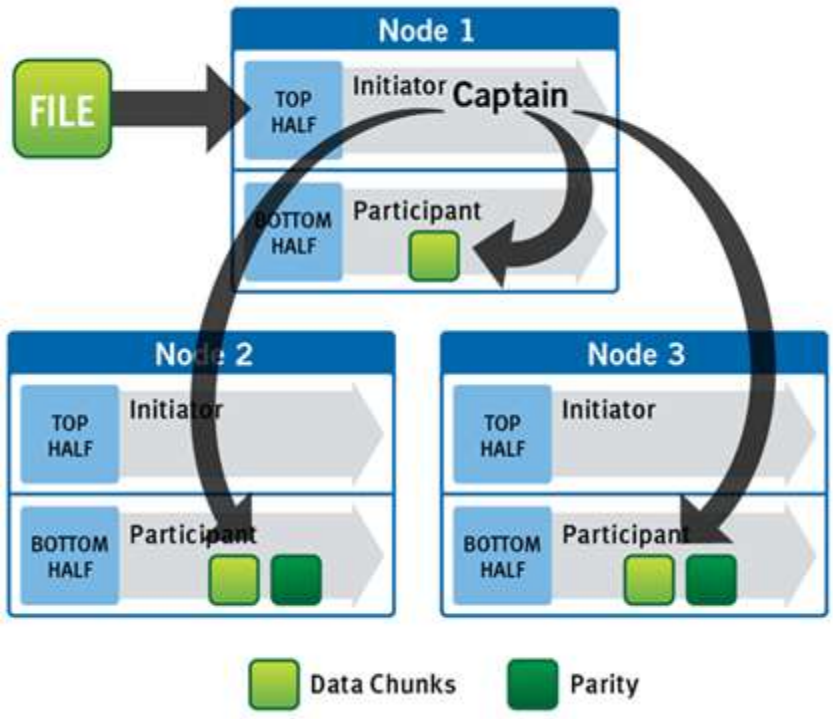

Figura 7: uma operação de gravação de arquivos em um cluster de três nós

O OneFS usa a rede de back-end para alocar e fracionar os dados em todos os nós do cluster automaticamente, assim, não é necessário nenhum processamento adicional. Como os dados estão sendo gravados, eles estão sendo protegidos no nível especificado. Quando a gravação ocorre, o OneFS divide os dados em unidades atômicas chamadas de grupos de proteção. A redundância é integrada a grupos de proteção de modo que, se cada grupo de proteção estiver protegido, todo o arquivo está protegido. Para arquivos protegidos pelos códigos de eliminação, um grupo de proteção consiste em uma série de blocos de dados, bem como um conjunto de códigos de eliminação desses blocos de dados; para os arquivos espelhados, um grupo de proteção consiste em todos os espelhos de um conjunto de blocos. O OneFS é capaz de alternar o tipo de grupo de proteção utilizado em um arquivo dinamicamente, à medida que está gravando. Isso pode permitir muitos recursos adicionais, incluindo, por exemplo, permitir que o sistema continue sem bloqueio em situações em que as falhas de nó temporárias no cluster evitariam que o número desejado de códigos de eliminação fosse usado. O espelhamento pode ser usado temporariamente nesses casos para permitir que as gravações continuem. Quando os nós são restaurados para o cluster, esses grupos de proteção espelhados são convertidos de volta perfeita e automática para serem protegidos por código contra eliminação, sem a intervenção do administrador.

O tamanho de block do file system do OneFS é 8 KB. Um arquivo menor do que 8 KB usará um block de 8 KB completo. Dependendo do nível de proteção de dados, esse arquivo de 8 KB pode acabar usando mais de 8 KB de espaço de dados. No entanto, as configurações de proteção de dados são discutidas em detalhes em uma seção posterior deste documento. O OneFS pode aceitar file systems com bilhões de arquivos pequenos com alto desempenho, pois todas as estruturas em disco são projetadas para serem dimensionadas para esses tamanhos e fornecer acesso quase instantâneo a qualquer objeto, independentemente do número total de objetos. Para arquivos maiores, o OneFS pode tirar vantagem de utilizar múltiplos blocks de 8 KB contíguos. Nesses casos, até 16 blocks contíguos podem ser fracionados em um só disco do nó. Se um arquivo tiver 32 KB, os quatro blocks de 8 KB contíguos serão utilizados.

Para arquivos ainda maiores, o OneFS pode maximizar o desempenho sequencial, aproveitando uma unidade de fração composta por 16 blocks contíguos para um total de 128 KB por unidade de fração. Durante a gravação, os dados são divididos em unidades de fração, e elas são espalhadas em vários nós como um grupo de proteção. Como os dados estão sendo exibidos em todo o cluster, os códigos de eliminação ou os mirrors, conforme necessário, são distribuídos dentro de cada grupo de proteção para garantir que os arquivos estejam sempre protegidos.

Uma das funções principais da funcionalidade AutoBalance do OneFS é realocar e reequilibrar dados, além de tornar o espaço de armazenamento mais útil e eficiente, quando possível. Na maioria dos casos, a largura da fração de arquivos maiores pode ser aumentada para aproveitar o novo espaço livre (conforme são adicionados nós) e para tornar o fracionamento em disco mais eficiente. O AutoBalance mantém alta eficiência em disco e elimina os pontos de acesso em disco automaticamente.

A metade superior de iniciador do nó “capitão” utiliza uma transação de confirmação modificada de duas fases para distribuir com segurança as gravações em várias NVRAMs em todo o cluster, como mostrado na Figura 8 abaixo.

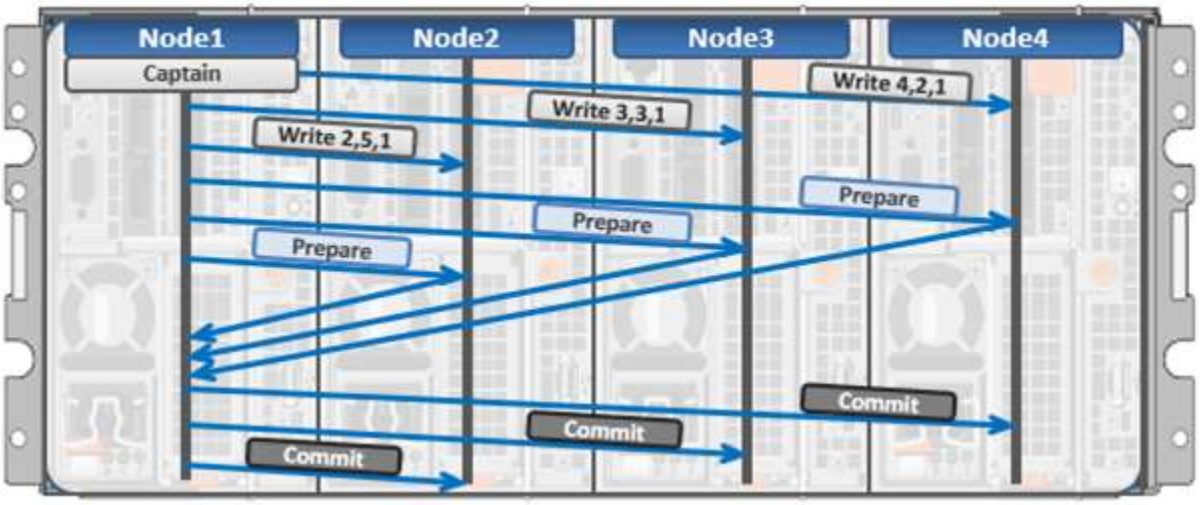

Figura 8: transações distribuídas e confirmação de duas fases

Cada nó que possui blocks em uma gravação específica está envolvido em uma confirmação de duas fases. O mecanismo baseia-se em NVRAM para registrar todas as transações que estão ocorrendo em cada nó no cluster de armazenamento. Usar várias NVRAMs em paralelo permite gravações com alto throughput, mantendo a segurança dos dados contra todos os tipos de falhas, incluindo falhas de energia. Em caso de falha em um nó durante a transação, a transação é reiniciada instantaneamente sem que o nó seja envolvido. Quando o nó retorna, as únicas ações necessárias são: o nó reproduzir seu registro a partir da NVRAM – o que leva alguns segundos ou minutos – e, ocasionalmente, o AutoBalance refazer o balanceamento dos arquivos envolvidos na transação. Nenhum processo caro de “fsck” ou “verificação de disco” é necessário. Nenhuma ressincronização prolongada precisa ser realizada. As gravações nunca são bloqueadas devido a uma falha. Este sistema de transação patenteada é uma das formas de o OneFS eliminar pontos únicos – e até mesmo vários – de falha.

Em uma operação de gravação, o iniciador “comanda” ou organiza o layout dos dados e metadados, a criação de códigos de eliminação e as operações normais de gerenciamento de bloqueio e controle de permissões. Um administrador de gerenciamento da Web ou interface CLI, a qualquer momento, pode otimizar decisões de layout tomadas pelo OneFS para melhor adequar-se ao workflow. O administrador pode escolher dentre os padrões de acesso abaixo em nível de arquivo ou diretório:

- Simultaneidade: otimiza a carga atual no cluster, com diversos clients simultâneos. Essa configuração oferece o melhor comportamento para cargas de trabalho mistas.
- Streaming: otimiza o streaming de alta velocidade de um arquivo único, por exemplo, para permitir leitura muito rápida com um só client.
- Aleatório: otimiza o acesso imprevisível ao arquivo, ajustando o fracionamento e desabilitando o uso de qualquer cache de pré-busca.

O OneFS também inclui pré-busca adaptável em tempo real, proporcionando o desempenho ideal de leitura para arquivos com um padrão de acesso reconhecível, sem nenhuma intervenção administrativa.

ⓘ O maior tamanho de arquivo compatível atualmente com o OneFS aumentou para 16 TB no OneFS 8.2.2 e posterior, a partir de um máximo de 4 TB em versões anteriores.

Armazenamento em cache do OneFS

O projeto de infraestrutura de armazenamento em cache do OneFS é estabelecido ao agregar o cache presente em cada nó de um cluster em um pool de memória globalmente acessível. Para fazer isso, o OneFS usa um sistema de mensagens eficiente, semelhante ao NUMA (Non-Uniform Memory Access, acesso a memória não uniforme). Isso permite que todo o cache de memória dos nós esteja disponível para todos os nós do cluster. A memória remota é acessada por meio de uma interconexão interna e tem latência muito menor do que o acesso às unidades de disco rígido.

Para o acesso remoto à memória, o OneFS utiliza uma rede Ethernet simples, redundante e subscrita essencialmente como um barramento de sistema distribuído. Embora não seja tão rápido quanto a memória local, o acesso remoto à memória ainda é muito rápido devido à baixa latência da Ethernet de 40 Gb.

O subsistema de armazenamento em cache do OneFS é coerente em todo o cluster. Isso significa que, se o mesmo conteúdo existir nos caches privados de vários nós, os dados armazenados em cache serão consistentes em todas as instâncias. O OneFS utiliza o protocolo MESI para manter a coerência do cache. Esse protocolo implementa uma política de “invalidação em gravação” para garantir que todos os dados sejam consistentes em todo o cache compartilhado.

O OneFS usa até três níveis de cache de leitura, além de um cache de gravação com backup NVRAM ou um aglutinador. Esses, e sua interação de alto nível, são ilustrados no diagrama a seguir.

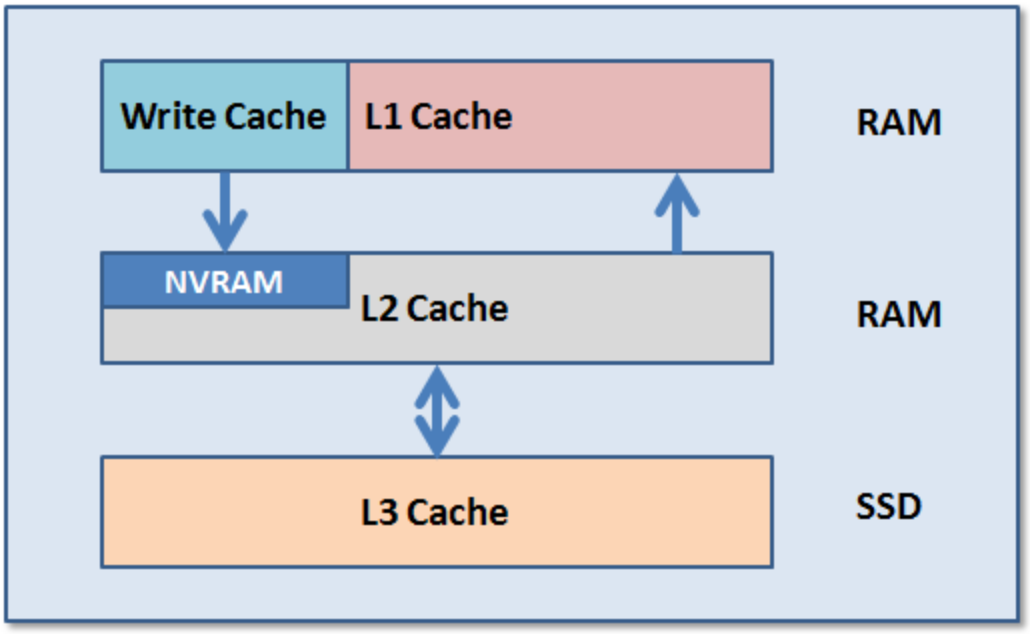

Figura 9: hierarquia de armazenamento em cache do OneFS

Os primeiros dois tipos de cache de leitura, nível 1 (L1) e nível 2 (L2), são baseados em memória (RAM) e são análogos ao cache usado em processadores (CPUs). Essas duas camadas de cache estão presentes em todos os nós de armazenamento da plataforma.

Nome	Type	Persistência	Descrição
Cache L1	RAM	Volátil	Também chamado de cache de front-end, mantém cópias limpas e coerentes com o cluster de dados do file system e blocks de metadados solicitados por clients pela rede de front-end
Cache L2	RAM	Volátil	Cache de back-end, que contém cópias limpas dos dados e metadados do file system em um nó local
SmartCache / Aglutinador de gravação	NVRAM	Não volátil	Cache de registro NVRAM persistente com bateria reserva que armazena em buffer todas as gravações pendentes em arquivos de front-end que não foram confirmados no disco.
SmartFlash Cache L3	SSD	Não volátil	Contém os dados em arquivo e blocos de metadados excluídos do cache L2, aumentando efetivamente a capacidade do cache L2.

Coerência de cache do OneFS

O subsistema de armazenamento em cache do OneFS é coerente em todo o cluster. Isso significa que, se o mesmo conteúdo existir nos caches privados de vários nós, os dados armazenados em cache serão consistentes em todas as instâncias. Por exemplo, considere o estado inicial e a sequência de eventos a seguir:

1. O nó 1 e o nó 5 têm uma cópia dos dados localizados em um endereço no cache compartilhado.
2. O nó 5, em resposta a uma solicitação de gravação, invalida a cópia do nó 1.
3. Em seguida, o nó 5 atualiza o valor. (Consulte abaixo).
4. O nó 1 deve ler novamente os dados do cache compartilhado para obter o valor atualizado.

O OneFS utiliza o protocolo MESI para manter a coerência do cache. Esse protocolo implementa uma política de "invalidação em gravação" para garantir que todos os dados sejam consistentes em todo o cache compartilhado. O diagrama a seguir ilustra os vários estados que os dados no cache podem ter, bem como as transições entre eles. Os vários estados na figura são:

- M – Modificado: os dados existem somente no cache local e foram alterados a partir do valor no cache compartilhado. Dados modificados são normalmente chamados de sujos.
- E – Exclusivo: os dados existem somente no cache local, mas correspondem ao que está no cache compartilhado. Esses dados muitas vezes são chamados de limpos.
- S – Compartilhado: os dados no cache local também podem estar em outros caches locais do cluster.
- I – Inválido: um bloqueio (exclusivo ou compartilhado) foi perdido nos dados.

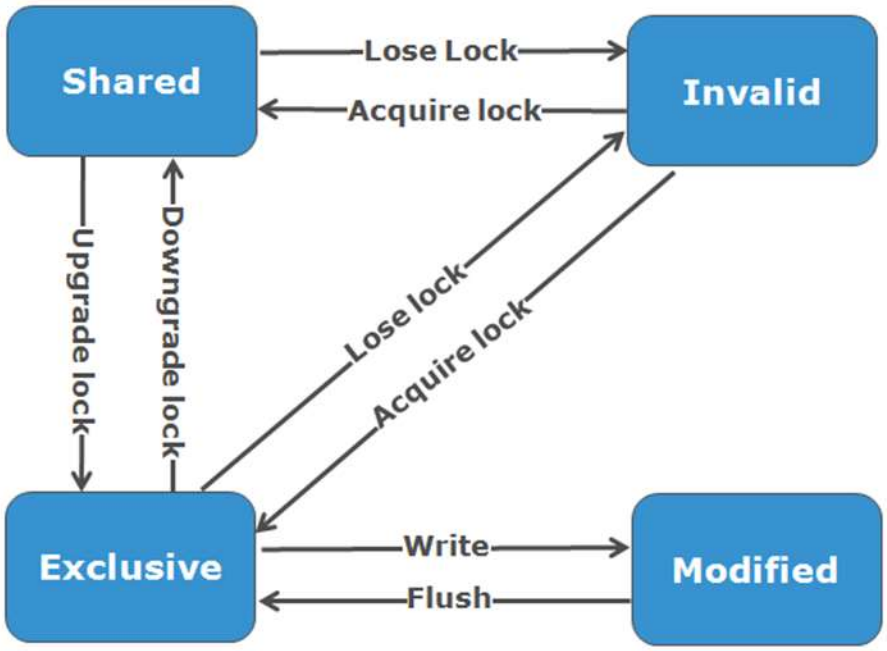

Figura 10: diagrama de estado de coerência de cache do OneFS

Cache de nível 1

O cache de nível 1 (L1), ou cache de front-end, é a memória mais próxima das camadas de protocolo (por exemplo, NFS, SMB etc.) usada pelos clients ou iniciadores, conectados a esse nó. O objetivo principal do cache L1 é fazer pré-busca dos dados de nós remotos. Os dados têm pré-busca por arquivo, e isso é otimizado para reduzir a latência associada à rede de back-end dos nós. Como a latência de interconexão de back-end é relativamente pequena, o tamanho do cache L1 e o volume típico de dados armazenados por solicitação, é menor do que o cache L2.

O L1 também é conhecido como cache remoto, pois contém dados recuperados de outros nós no cluster. Ele é coerente em todo o cluster, mas é usado apenas pelo nó no qual reside e não pode ser acessado por outros nós. Os dados no cache L1 nos nós de armazenamento são descartados de maneira agressiva depois de utilizados. O cache L1 usa endereçamento baseado em arquivo, no qual os dados são acessados por meio de um deslocamento para um objeto de arquivo.

O cache L1 refere-se à memória no mesmo nó do iniciador. Ele só pode ser acessado para o nó local e, normalmente, o cache não é a cópia principal dos dados. Isso é análogo ao cache L1 em um núcleo de CPU, que pode ser invalidado conforme outros núcleos gravam na memória principal.

A coerência de cache L1 é gerenciada por meio de um protocolo tipo MESI que usa bloqueios distribuídos, conforme descrito acima.

O OneFS também usa um cache de inodes dedicado no qual os inodes recentemente solicitados são mantidos. O cache de inodes com frequência tem um grande impacto no desempenho, pois os clients geralmente armazenam dados em cache e muitas atividades de E/S de rede são principalmente solicitações de atributos de arquivos e metadados, que podem ser retornados rapidamente do inode armazenado em cache.

ⓘ O cache L1 é utilizado de maneira diferente nos nós aceleradores do cluster, que não contêm unidades de disco. Em vez disso, o cache de leitura inteiro é o cache L1, já que todos os dados são buscados de outros nós de armazenamento. Além disso, o amadurecimento (“aging”) do cache baseia-se em uma política de LRU (Least Recently Used, remoção menos recente), em oposição ao algoritmo de desativação usado normalmente no cache L1 de um nó de armazenamento. Como o cache L1 do acelerador é grande, e os dados nele têm muito mais chances de serem solicitados novamente, os blocos de dados não são removidos imediatamente do cache após o uso. No entanto, os metadados e as pesadas cargas de trabalho de atualização não se beneficiam tanto, e o cache de um acelerador é benéfico apenas aos clients conectados diretamente ao nó.

Cache de nível 2

O cache de nível 2 (L2), ou cache de back-end, refere-se à memória local no nó em que um block específico de dados é armazenado. O cache L2 pode ser acessado globalmente a partir de qualquer nó do cluster e é usado para reduzir a latência de uma operação de leitura, não exigindo uma busca diretamente das unidades de disco. Dessa forma, o volume de dados com pré-busca no cache L2 para ser usado por nós remotos é muito maior do que no cache L1.

O cache L2 também é conhecido como cache local, pois contém dados recuperados de unidades de disco localizadas nesse nó e, em seguida, disponibilizados para solicitações de nós remotos. Os dados no cache L2 são removidos de acordo com um algoritmo de LRU.

Os dados no cache L2 são resolvidos pelo nó local usando uma deslocamento em uma unidade de disco que é local para esse nó. Como o nó sabe onde os dados solicitados pelos nós remotos estão localizados no disco, esse é um modo muito rápido de recuperar dados destinados a nós remotos. Um nó remoto acessa o cache L2 fazendo uma pesquisa do endereço de block de um objeto de arquivo específico. Conforme descrito acima, não é necessária nenhuma invalidação de MESI aqui, além disso, o cache é atualizado automaticamente durante as gravações e mantido coerente pelo sistema de transações e NVRAM.

Cache de nível 3

Um terceiro nível opcional de cache de leitura, chamado de SmartFlash ou cache de nível 3 (L3), também é configurável nos nós que contêm SSDs (Solid State Drives, drive de estado sólido). O SmartFlash (L3) é um cache de exclusão preenchido por blocks de cache L2, pois são excluídos da memória por serem antigos. Há diversos benefícios proporcionados pelo uso de SSDs para armazenamento em cache em vez de dispositivos tradicionais de armazenamento em file system. Por exemplo, quando reservado para o armazenamento em cache, todo o SSD será usado, e as gravações ocorrerão de maneira muito linear e previsível. Isso garante uma utilização muito melhor e também resulta em um desgaste muito menor e maior durabilidade no uso normal do file system, especialmente com cargas de trabalho de gravação aleatórias. O SSD para cache também torna a capacidade de dimensionamento do SSD muito mais simples e menos propenso a erros em comparação ao uso de SSDs como um nível de armazenamento.

O diagrama a seguir ilustra como os clients interagem com a infraestrutura do cache de leitura do OneFS e o aglutinador de gravação. O cache L1 ainda interage com o cache L2 em qualquer nó exigido, e o cache L2 interage com o subsistema de armazenamento e o cache L3. O cache L3 é armazenado em um SSD dentro do nó e cada nó no mesmo pool de nós tem o cache L3 habilitado.

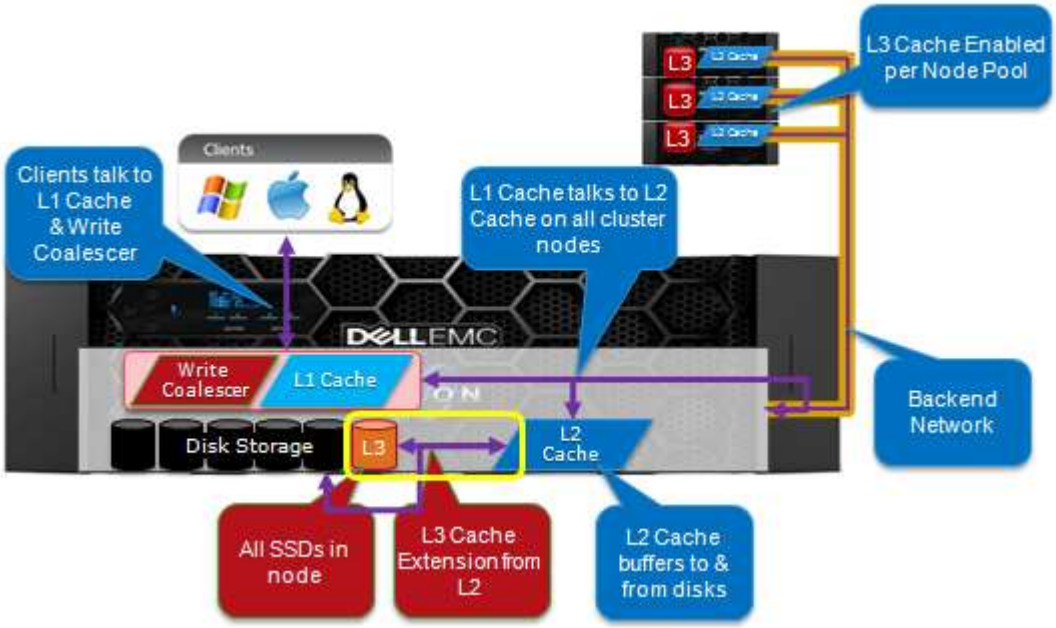

Figura 11: arquitetura de armazenamento em cache L1, L2 e L3 do OneFS

O OneFS determina que um arquivo seja gravado em vários nós no cluster e, possivelmente, em várias unidades dentro de um nó, para que todas as solicitações de leitura envolvam a leitura de dados remotos (e possivelmente locais). Quando uma solicitação de leitura chega de um client, o OneFS determina se os dados solicitados estão no cache local. Todos os dados residentes no cache local são imediatamente lidos. Se os dados solicitados não estão no cache local, eles são lidos no disco. Para os dados que não estão no nó local, uma solicitação é feita a partir dos nós remotos nos quais eles residem. Em cada um dos outros nós, outra pesquisa de cache é realizada. Todos os dados no cache são retornados imediatamente, e todos os dados que não estão no cache são recuperados do disco.

Quando os dados são recuperados do cache local e remoto (e possivelmente do disco), eles são devolvidos ao client.

As etapas de alto nível para atender a uma solicitação de leitura em um nó local e remoto são:

No nó local (o nó que recebe a solicitação):

1. Determine se parte dos dados solicitados está no cache L1 local. Em caso afirmativo, retorne ao client.
2. Se não estiver no cache local, solicite os dados do(s) nó(s) remoto(s).

Em nós remotos:

1. Determine se parte dos dados solicitados está no cache L2 ou L3 local. Em caso afirmativo, retorne para o nó solicitante.
2. Se não estiver no cache local, leia a partir do disco e retorne para o nó solicitante.

O armazenamento em cache para gravação acelera o processo de gravação de dados ao cluster. Isso é obtido criando-se lotes de solicitações de gravação menores e enviando-os para o disco em fragmentos maiores, removendo uma quantidade significativa de latência de gravação em disco. Quando os clients gravam no cluster, o OneFS temporariamente grava os dados em um cache de registro baseado em NVRAM no nó iniciador, em vez de imediatamente gravar no disco. O OneFS pode fazer flush dessas gravações em cache no disco posteriormente, quando for mais conveniente. Além disso, essas gravações também são espelhadas para registros de NVRAM dos nós participantes para atender aos requisitos de proteção do arquivo. Portanto, no caso de uma divisão do cluster ou paralisação inesperada do nó, as gravações em cache não confirmadas estão totalmente protegidas.

O cache de gravação funciona da seguinte forma:

- Um client NFS envia ao Nó 1 uma solicitação de gravação para um arquivo com proteção +2n.
- O Nó 1 aceita as gravações em seu cache de gravação NVRAM (caminho rápido) e depois espelha as gravações em arquivos de registro do nó participante para fins de proteção.

- Os reconhecimentos de gravação são devolvidos ao client NFS imediatamente e, desse modo, a latência de gravação em disco é evitada.
- À medida que o cache de gravação do Nó 1 é preenchido, ele recebe flush periodicamente e as gravações são confirmadas no disco através do processo de confirmação de duas fases (descrito acima) com a proteção de ECC (erasure code, código de eliminação) adequada aplicada (+2n).
- Os arquivos de registro do nó participante e do cache de gravação são removidos e ficam disponíveis para aceitar novas gravações.

📖 Mais informações estão disponíveis no white paper OneFS SmartFlash.

Leituras de arquivo

Dados, metadados e inodes estão distribuídos em vários nós e até mesmo em várias unidades dentro de nós. Ao ler ou gravar no cluster, o nó ao qual um client se conecta atua como o "capitão" da operação.

Em uma operação de leitura, o nó "capitão" reúne todos os dados dos vários nós do cluster e os apresenta de forma coesa ao solicitante.

Devido ao uso de hardware com custo otimizado padrão do setor, o cluster fornece uma alta taxa de cache em disco (vários GB por nó), que é alocado dinamicamente para operações de leitura e gravação quando necessário. Esse cache baseado em RAM é unificado e coerente em todos os nós do cluster, permitindo que uma solicitação de leitura do client em um nó se beneficie de E/S já realizada em outro nó. Esses Blocks armazenados em cache podem ser acessados rapidamente a partir de qualquer nó no backplane de baixa latência, permitindo um cache de RAM grande e eficiente, o que acelera muito o desempenho de leitura.

À medida que o cluster aumenta, os benefícios do cache também aumentam. Por essa razão, o volume de E/S em disco em um cluster geralmente é substancialmente inferior ao que é em plataformas tradicionais, permitindo latências reduzidas e uma melhor experiência do usuário.

Para arquivos marcados com um padrão de acesso de simultâneo ou de streaming, o OneFS pode aproveitar a pré-busca de dados com base em heurística utilizada pelo componente SmartRead. O SmartRead pode criar um "pipeline" de dados do cache L2, fazendo pré-busca em um cache "L1" local sobre o nó "capitão". Isso melhora muito o desempenho de leitura sequencial em todos os protocolos e significa que as leituras vêm diretamente da RAM em milésimos de segundos. Para casos altamente sequenciais, o SmartRead pode fazer pré-busca adiantada de maneira muito agressiva, permitindo leituras e gravações de arquivos individuais com taxas de dados muito altas.

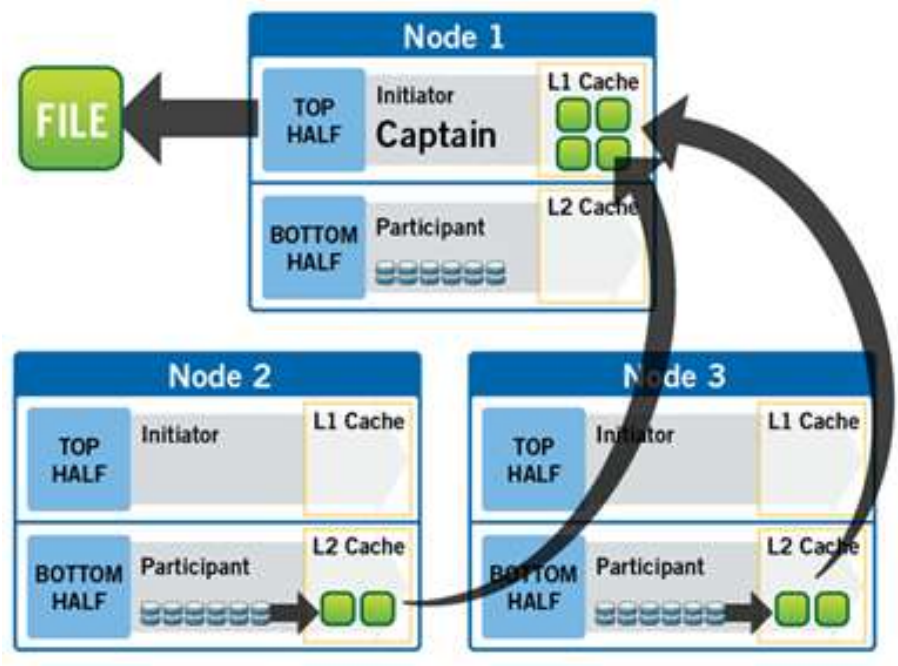

Figura 12: Uma operação de leitura de arquivos em um cluster de três nós

A Figura 10 ilustra como o SmartRead lê um arquivo sequencialmente acessado que não está em cache e que é solicitado por um client conectado ao Nó 1 em um cluster de 3 nós.

1. O Nó 1 lê os metadados para identificar onde estão todos os blocos de dados em arquivo.
2. O Nó 1 também verifica seu cache L1 para ver se ele tem os dados em arquivo que estão sendo solicitados.
3. O Nó 1 cria um pipeline de leitura, enviando solicitações simultâneas a todos os nós que têm algum dado de arquivo para recuperar esses dados do disco.
4. Cada nó recebe os blocos de dados em arquivo do disco para seu cache L2 (ou cache L3 SmartFlash, quando disponível) e transmite os dados em arquivo para o Nó 1.
5. O Nó 1 registra os dados recebidos no cache L1 ao mesmo tempo que entrega o arquivo para o client. Enquanto isso, o processo de pré-busca continua.
6. Para casos altamente sequenciais, os dados no cache L1 podem ser opcionalmente “deixados para trás” para liberar a memória RAM para outras demandas de cache L1 ou L2.

O cache inteligente do SmartRead permite desempenho de leitura muito alto com altos níveis de acesso simultâneo. E o mais importante, é mais rápido para o Nó 1 obter dados em arquivo do cache do Nó 2 (sobre a interconexão do cluster de baixa latência) do que acessar seu próprio disco local. Os algoritmos do SmartRead controlam o nível de rigor de pré-busca (desativando a pré-busca para casos de acesso aleatório) e quanto tempo os dados permanecem no cache além de otimizar onde os dados são armazenados em cache.

Bloqueios e simultaneidade

O OneFS tem um gerenciador de bloqueio totalmente distribuído que organiza bloqueios em dados em todos os nós em um cluster de armazenamento. O gerenciador de bloqueio é altamente extensível e permite múltiplas “personalidades” de bloqueio para dar suporte a bloqueios do file system e a bloqueios em nível de protocolo coerentes com o cluster como bloqueios em modo de compartilhamento de SMB ou em modo de recomendação do NFS. O OneFS também tem suporte para bloqueios delegados, como bloqueios CIFS e delegações NFSv4.

Cada nó em um cluster é um coordenador de recursos de bloqueio, e um coordenador é atribuído a recursos bloqueáveis com base em um algoritmo de hash avançado. A forma como o algoritmo é projetado é que o coordenador quase sempre termina em um nó diferente do que o iniciador da solicitação. Quando é solicitado um bloqueio de um arquivo, ele pode ser um bloqueio compartilhado (permitindo que múltiplos usuários compartilhem o bloqueio simultaneamente, geralmente para leituras) ou um bloqueio exclusivo (permitindo um usuário em um dado momento, tipicamente para gravações).

A Figura 13 abaixo ilustra um exemplo de como threads de diferentes nós poderiam solicitar um bloqueio a partir do coordenador.

1. O Nó 2 é designado para ser o coordenador desses recursos.
2. O thread 1 do Nó 4 e o thread 2 do Nó 3 solicitam um bloqueio compartilhado em um arquivo do Nó 2 ao mesmo tempo.
3. O Nó 2 verifica se existe um bloqueio exclusivo para o arquivo solicitado.
4. Se não existirem bloqueios exclusivos, o Nó 2 concederá ao thread 1 do Nó 4 e ao thread 2 do Nó 3 bloqueios compartilhados do arquivo solicitado.
5. O Nó 3 e o Nó 4 estão agora realizando uma leitura com base no arquivo solicitado.
6. O thread 3 do Nó 1 solicita um bloqueio exclusivo do mesmo arquivo que está sendo lido pelo Nó 3 e pelo Nó 4.
7. O Nó 2 verifica com o Nó 3 e o Nó 4 se os bloqueios compartilhados podem ser recuperados.
8. O Nó 3 e o nó 4 ainda estão fazendo a leitura, portanto, o Nó 2 solicita ao thread 3 do Nó 1 que aguarde um instante.
9. O thread 3 do Nó 1 é bloqueado até que o bloqueio exclusivo seja concedido pelo Nó 2 e, em seguida, conclui a operação de gravação.

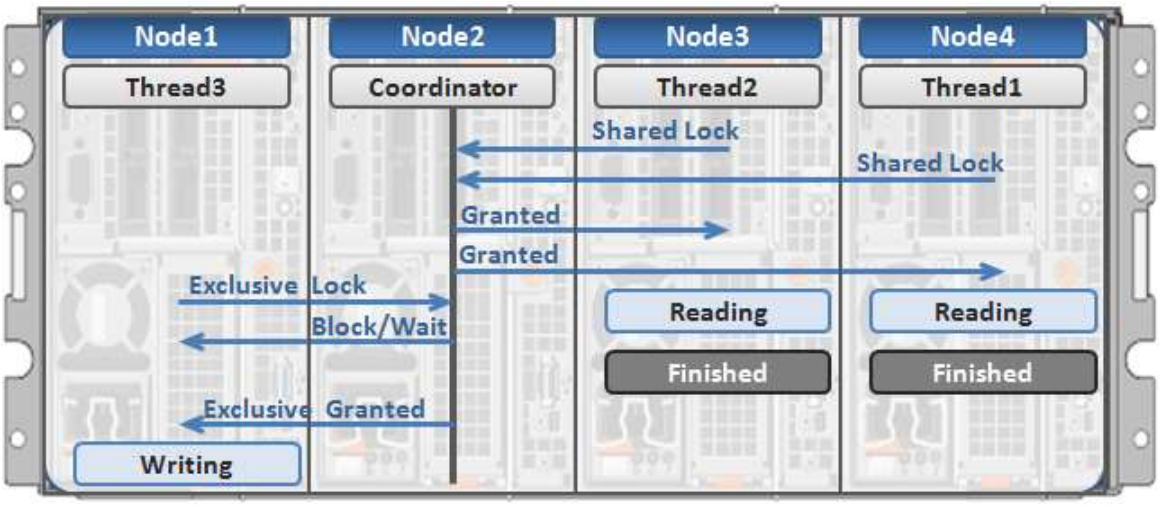

Figura 13: gerenciador de bloqueio distribuído

E/S com múltiplos threads

Com o crescente uso de grandes datastores NFS para virtualização de servidores e suporte a aplicativos corporativos, surge a necessidade de alto throughput e baixa latência para arquivos grandes. Para acomodar isso, o OneFS Multi-writer dá suporte a gravação de múltiplos threads simultaneamente para arquivos individuais.

No exemplo acima, o acesso simultâneo de gravação a um arquivo grande pode ficar limitado pelo mecanismo de bloqueio exclusivo, aplicado no nível do arquivo todo. A fim de evitar esse possível gargalo, o OneFS Multi-writer fornece bloqueio de gravação mais granular subdividindo o arquivo em regiões separadas e concedendo bloqueios de gravação exclusivos para cada região, em vez de todo o arquivo. Desse modo, vários clients podem gravar simultaneamente em partes diferentes do mesmo arquivo.

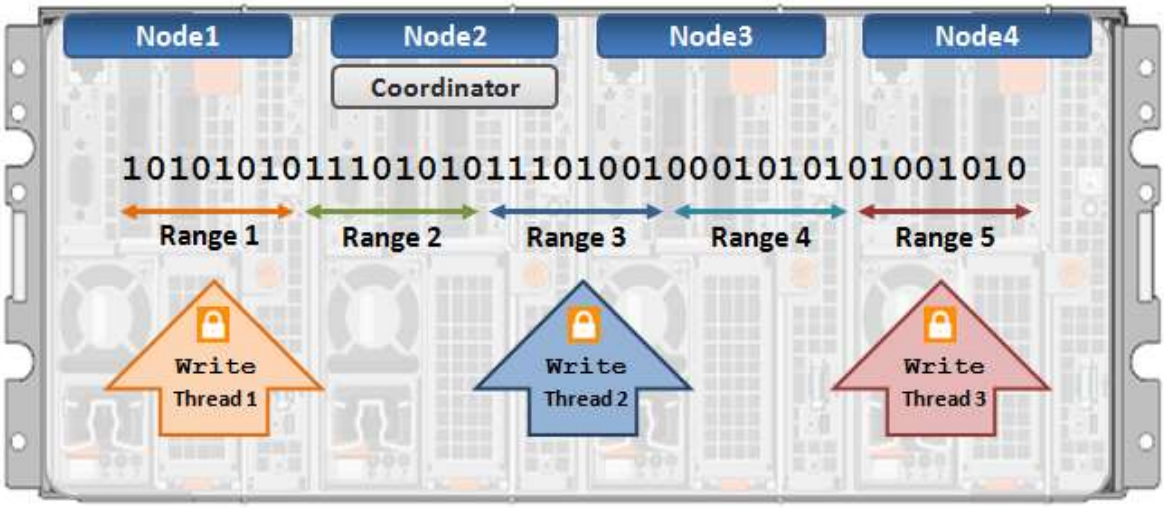

Figura 14: Gravador de E/S com múltiplos threads

Proteção de dados

Perda de alimentação

Um registro do file system, que armazena informações sobre as alterações no file system, é projetado para permitir rápidas recuperações consistentes após falhas do sistema ou outras falhas, como perda de alimentação. O file system reproduz as entradas do registro depois que um nó ou cluster se recupera de uma perda de alimentação ou outra suspensão temporária de força. Sem um registro, um file system precisaria examinar e analisar cada possível alteração individualmente após uma falha (uma operação “fsck” ou “chkdsk”). Em um file system grande, essa operação pode levar muito tempo.

O OneFS é um file system com registro no qual cada nó contém um cartão NVRAM com bateria reserva usada para proteger gravações não confirmadas no file system. A carga da bateria NVRAM dura muitos dias sem a necessidade de recarga. Quando um nó é inicializado, ele verifica seu registro e reproduz as transações seletivamente no disco onde o sistema de registro considerar necessário.

O OneFS vai ser montado só se puder garantir que todas as transações que ainda não estão no sistema foram registradas. Por exemplo, se os procedimentos de desligamento apropriados não foram seguidos e a bateria NVRAM foi descarregada, as transações poderão ter sido perdidas. Para evitar possíveis problemas, o nó não montará o file system.

Falhas de hardware e quórum

Para que o cluster funcione corretamente e aceite gravações de dados, um quórum de nós deve estar ativo e respondendo. Um quórum é definido como uma maioria simples: um cluster com nós deve ter ⌊/2⌋+1 nós on-line para permitir gravações. Por exemplo, em um cluster de sete nós, quatro nós seriam necessários para um quórum. Se um nó ou grupo de nós está funcionando e é responsivo, mas não é membro de um quórum, ele é executado em um estado somente leitura.

O OneFS usa um quórum para evitar condições de “síndrome de desconexão” que podem ser introduzidas se o cluster for temporariamente dividido em dois clusters. Seguindo a regra de quórum, a arquitetura garante que, independentemente de quantos nós falharem ou voltarem a ficar on-line, se uma gravação ocorrer, poderá ficar consistente com qualquer gravação anterior que tenha ocorrido. O quórum também determina o número de nós necessários para mover-se para um determinado nível de proteção de dados. Para um nível de proteção baseada em código de eliminação de +, o cluster deve conter pelo menos 2+1 2+1 nós. Por exemplo, um mínimo de sete nós é necessário para uma configuração de +3n, o que permite uma perda simultânea de três nós, e ao mesmo tempo mantém um quórum de quatro nós para o cluster permanecer operacional. Se um cluster cair abaixo do quórum, o file system será automaticamente colocado em um estado protegido e somente leitura, negando gravações, mas ainda permitindo o acesso de leitura aos dados disponíveis.

Falhas de hardware – adicionar/remover nós

Um sistema chamado de GMP (group management protocol, protocolo de gerenciamento de grupo) permite o conhecimento global do estado do cluster em todos os momentos e garante uma visão consistente, em todo o cluster, do estado de todos os outros nós. Se um ou mais nós ficarem inatingíveis sobre a interconexão do cluster, o grupo será “dividido” ou removido do cluster. Todos os nós são resolvidos para uma nova exibição coerente de seu cluster. (Considere isso como se o cluster estivesse se dividindo em dois grupos separados de nós, mas observe que apenas um grupo pode ter quórum.) No estado de divisão, todos os dados no file system podem ser acessados e, para a porção de manutenção do quórum, modificáveis. Todos os dados armazenados no dispositivo “de baixo” são reconstruídos usando a redundância armazenada no cluster.

Se o nó fica novamente acessível, ocorre uma “mesclagem” ou adição, trazendo o(s) nó(s) de volta para o cluster. (Os dois grupos se mesclam novamente em um.) O nó pode reingressar no cluster sem ser reconstruído e reconfigurado. Isso é diferente de arrays RAID de hardware, que exigem que unidades sejam reconstruídas. O AutoBalance pode fracionar novamente alguns arquivos para aumentar a eficiência se alguns de seus grupos de proteção foram substituídos e transformados em frações mais restritas de proteção durante a divisão.

O OneFS também inclui um processo chamado de Collect, que atua como um colecionador de órfãos. Quando um cluster é dividido durante uma operação de gravação, alguns blocks que foram alocados para o arquivo podem ter de ser realocados na porção do quórum. Isso irá deixar “órfãos” os blocks alocados na porção não quórum. Quando o cluster é mesclado novamente, o trabalho Collect localiza esses blocks órfãos através de uma verificação do tipo marcar e limpar paralelizada e recupera-os como espaço livre para o cluster.

Reconstrução escalável

O OneFS não depende de RAID de hardware para alocação de dados nem para a reconstrução dos dados após as falhas. Em vez disso, o OneFS gerencia a proteção de dados em arquivo diretamente, e, quando ocorre uma falha, ele reconstrói os dados de uma forma paralelizada. O OneFS é capaz de determinar quais arquivos são afetados por uma falha em tempo constante, lendo os de dados do inode de maneira linear diretamente do disco. O conjunto de arquivos afetados é atribuído a um conjunto de threads de operador que são distribuídos entre os nós do cluster pelo Job Engine. Os nós do operador reparam os arquivos em paralelo. Isso implica que, à medida que o tamanho do cluster aumenta, o tempo de reconstrução a partir de falhas diminui. Isso tem uma vantagem enorme em termos de eficiência na manutenção da capacidade de recuperação dos clusters à medida que seu tamanho aumenta.

Hot spare virtual

Sistemas de armazenamento mais tradicionais baseados em RAID exigem o provisionamento de uma ou mais unidades de “hot spare” para permitir a recuperação independente de unidades com falha. A unidade de hot spare substitui a unidade com falha em um conjunto de RAID. Se esses hot spares não são substituídos antes que apareçam mais falhas, o sistema corre o risco de uma perda de dados catastrófica. O OneFS evita a utilização de unidades de hot spare e simplesmente pega emprestado a partir do espaço livre disponível no sistema a fim de se recuperar de falhas; essa técnica é chamada de hot spare virtual. Ao fazer isso, ele permite que o cluster seja totalmente de autocorreção, sem intervenção humana. O administrador pode criar uma reserva de hot spare virtual, permitindo que o sistema tenha autocorreção apesar das gravações contínuas por parte dos usuários.

Proteção de dados em nível de arquivo com codificação de eliminação

Um cluster é projetado para tolerar uma ou mais falhas de componentes simultâneos sem impedir o cluster de enviar dados. Para conseguir isso, o OneFS protege os arquivos com a proteção baseada em código de eliminação, por meio da correção de erros Reed-Solomon (proteção N+M) ou de um sistema de espelhamento. A proteção de dados é aplicada no software no nível de arquivo, permitindo que o sistema se concentre em recuperar apenas os arquivos que estão comprometidos por uma falha, em vez de ter de verificar e reparar um conjunto de arquivos ou volume inteiro. Os metadados e inodes do OneFS estão sempre protegidos por meio de espelhamento, em vez da codificação Reed-Solomon, e com, pelo menos, o nível de proteção dos dados a que eles fazem referência.

Como todos os dados, metadados e informações de proteção são distribuídos entre os nós do cluster, o cluster não requer uma unidade nem nó de paridade dedicado, nem um dispositivo ou conjunto de dispositivos dedicado para gerenciar metadados. Isso garante que nenhum nó possa se tornar um ponto único de falha. Todos os nós compartilham igualmente as tarefas a serem executadas, proporcionando perfeita simetria e balanceamento de carga em uma arquitetura ponto a ponto.

O OneFS fornece vários níveis de ajustes configuráveis de proteção de dados, que você pode modificar a qualquer momento, sem a necessidade de colocar o cluster ou o file system off-line.

Para um arquivo protegido com códigos de eliminação, dizemos que cada um de seus grupos de proteção está protegido em um nível de N+M/b, em que N>M e M>=b. Os valores N e M representam, respectivamente, o número de unidades utilizadas para os dados e os códigos de eliminação dentro do grupo de proteção. O valor de b refere-se ao número de frações de dados usadas para exibir esse grupo de proteção e é abordado abaixo. Um caso comum e de fácil compreensão é em que b = 1, implicando que um grupo de proteção incorpora: o número de N unidades de dados; o número de M unidades de redundância, armazenados em códigos de eliminação; e que o grupo de proteção deve ser exibido sobre exatamente uma fração através de um conjunto de nós. Isso permite que M membros do grupo de proteção apresentem falha simultaneamente e ainda forneçam 100% de disponibilidade de dados. Os M membros do código de eliminação são calculados a partir do 푁 membros de dados. A figura 13 abaixo mostra o caso de um grupo de proteção regular 4+2 (N=4, M=2, b=1).

Como o OneFS fraciona arquivos entre os nós, isso implica que os arquivos fracionados em N+M podem suportar falhas simultâneas de nós sem perda de disponibilidade. O OneFS, portanto, fornece capacidade de recuperação em qualquer tipo de falha, para uma unidade, um nó ou um componente dentro de um nó (por exemplo, um cartão). Além disso, um nó conta como uma só falha, independentemente do número ou do tipo de componentes que falham dentro dele. Portanto, se cinco unidades falharem em um nó, ele contará como apenas uma falha para efeitos de proteção N+M.

O OneFS pode fornecer exclusivamente um nível variável de M, até quatro, fornecendo proteção contra falha quádrupla. Isso vai muito além do nível máximo do RAID usado atualmente, que é a proteção contra falhas duplas do RAID-6. Como a confiabilidade do armazenamento aumenta geometricamente com essa quantidade de redundância, a proteção de +4n pode ser uma ordem de magnitude mais confiável do que o RAID de hardware tradicional. Essa proteção adicional significa que unidades SATA de grande capacidade, como unidades de 4 TB e 6 TB, podem ser adicionadas com confiança.

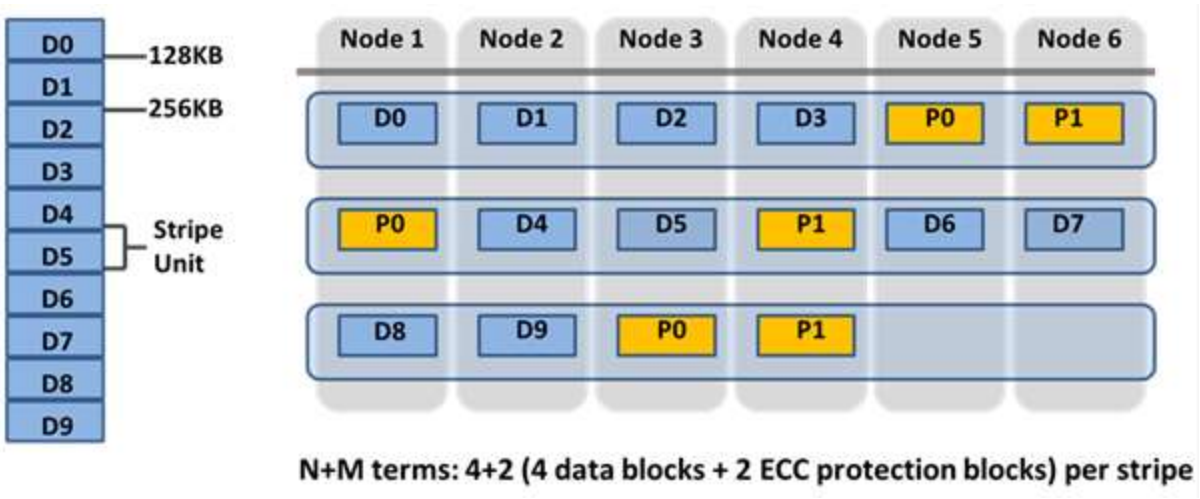

Figura 15: redundância do OneFS – proteção de código de eliminação N+M

Clusters menores podem ser protegidos com proteção +1n, mas isso implica que, enquanto uma só unidade ou nó podia ser recuperado, duas unidades para dois nós diferentes não podiam. As falhas em unidades são muito mais prováveis do que as falhas em nós. Para clusters com unidades de grande porte, é desejável fornecer proteção para falhas em várias unidades, porém, a capacidade de recuperação de um só nó é aceitável.

Para oferecer uma situação em que desejamos ter redundância de disco duplo e redundância de um único nó, podemos criar grupos de proteção de tamanho duplo ou triplo. Esses grupos de proteção de tamanho duplo ou triplo vão “cobrir” uma ou duas vezes o mesmo conjunto de nós à medida que são exibidos. Como cada grupo de proteção contém exatamente dois discos de redundância, esse mecanismo permitirá que um cluster mantenha uma falha de duas ou três unidades ou uma falha de nó completo, sem nenhuma indisponibilidade de dados.

O que é mais importante para clusters pequenos é que esse método de particionamento é altamente eficiente, com uma eficiência em disco de M/(N+M). Por exemplo, em um conjunto de cinco nós com proteção dupla contra falhas, estávamos usando N=3, M=2, e obtínhamos um grupo de proteção 3+2 com uma eficiência de 1−2/5 ou 60%. Usando o mesmo cluster de 5 nós, mas com cada grupo de proteção exibido sobre 2 frações, N agora seria 8 e M=2, portanto, poderíamos obter 1-2/(8+2) ou 80% de eficiência em disco, mantendo nossa proteção contra falha em dois discos e sacrificando somente a proteção contra falha em dois nós.

O OneFS aceita vários esquemas de proteção. Isso inclui o onipresente +2d:1n, que protege contra falhas de duas unidades ou falha de um nó.

ⓘ A prática recomendada é usar o nível de proteção recomendado para uma configuração particular em cluster. Esse nível de proteção recomendado é claramente marcado como “sugerido” nas páginas de configuração de pools de armazenamento da WebUI do OneFS e, em geral, é configurada por padrão. Para todas as configurações atuais de hardware Gen6, o nível de proteção recomendado é “+2d:1n”.

Os esquemas de proteção híbrida são particularmente úteis para o chassi de 6ª geração e configurações de nó de alta densidade, em que a probabilidade de várias unidades apresentarem falha supera de longe a de uma falha no nó inteiro. No caso improvável de vários dispositivos falharem ao mesmo tempo, de tal modo que o arquivo está “além de seu nível de proteção”, o OneFS protegerá novamente tudo o que for possível e relatará os erros nos arquivos individuais afetados nos registros do cluster.

O OneFS também disponibiliza uma variedade de opções de espelhamento variando de 2x a 8x, permitindo de 2 a 8 espelhamentos do conteúdo especificado. Os metadados, por exemplo, são espelhados em um nível acima da FEC por padrão. Por exemplo, se um arquivo é protegido em +2n, seu objeto associado de metadados será espelhado 3x.

A gama completa de níveis de proteção do OneFS está resumida na tabela a seguir:

Nível de proteção	Descrição
+1n	Tolerar a falha de uma unidade OU um nó
+2d:1n	Tolerar a falha de duas unidades OU um nó
+2n	Tolerar a falha de duas unidades OU dois nós
+3d:1n	Tolerar a falha de três unidades OU um nó
+3d:1n1d	Tolerar a falha de três unidades OU um nó E uma unidade
+3n	Tolerar a falha de três unidades OU três nós
+4d:1n	Tolerar a falha de quatro unidades OU um nó
+4d:2n	Tolerar a falha de quatro unidades OU dois nós
+4n	Tolerar a falha de quatro nós
2x a 8x	Espelhado em mais de dois a oito nós, dependendo da configuração

O OneFS permite que um administrador modifique a política de proteção em tempo real, enquanto os clients estão conectados e estão lendo e gravando dados.

ⓘ Observe que aumentar o nível de proteção de um cluster pode aumentar a quantidade de espaço consumido pelos dados no cluster.

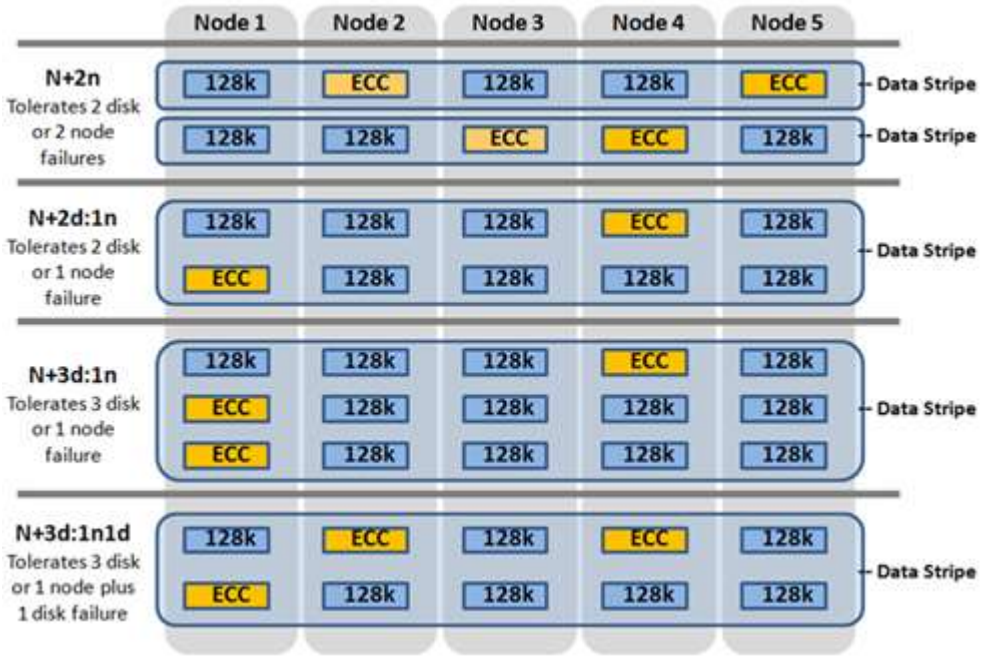

Figura 16: Esquemas de proteção híbrida do código de eliminação do OneFS

ⓘ O OneFS também oferece alertas de subproteção para novas instalações de cluster. Se o cluster subprotegido, o sistema CELOG (Cluster Event Logging, registro de eventos do cluster) gerará alertas, avisando o administrador da deficiência de proteção e recomendando uma alteração no nível de proteção apropriado para a configuração desse cluster específico.

🕮 Mais informações estão disponíveis no white paper Proteção de dados e alta disponibilidade do OneFS.

Particionamento automático

O gerenciamento e a classificação de dados por níveis no OneFS são manipulados pelo framework do SmartPools. De um ponto de vista de proteção de dados e eficiência de layout, o SmartPools facilita a subdivisão de grandes números de nós homogêneos de alta capacidade em pools de discos de menor MTTDL (Mean Time to Data Loss, tempo médio para perda de dados). Por exemplo, um cluster H500 near-line de 80 nós normalmente seria executado em um nível de proteção +3d:1n1d. No entanto, particionando-o em quatro, pools de discos de 20 nós permitiriam que cada pool fosse executado em +2d:1n, diminuindo assim a sobrecarga de proteção e melhorando a utilização do espaço sem nenhum aumento líquido na sobrecarga de gerenciamento.

Mantendo o objetivo de simplificar o gerenciamento de armazenamento, o OneFS calculará automaticamente e particionará o cluster em pools de discos, ou “pools de nós”, que são otimizados para a utilização eficiente de espaço e MTTDL. Isso significa que as decisões de nível de proteção, como o exemplo de cluster de 80 nós acima, não são deixadas para o cliente.

Com o provisionamento automático, cada conjunto de hardware de nó compatível é automaticamente dividido em pools de discos abrangendo até 40 nós e 6 unidades por nó. Esses pools de nó são protegidos por padrão em +2d:1n, e vários pools podem ser combinados em níveis lógicos e gerenciados com políticas de pool de arquivos do SmartPools. Pela subdivisão de discos de um nó em vários pools separadamente protegidos, os nós ficam significativamente mais resistentes a várias falhas de disco do que era possível anteriormente.

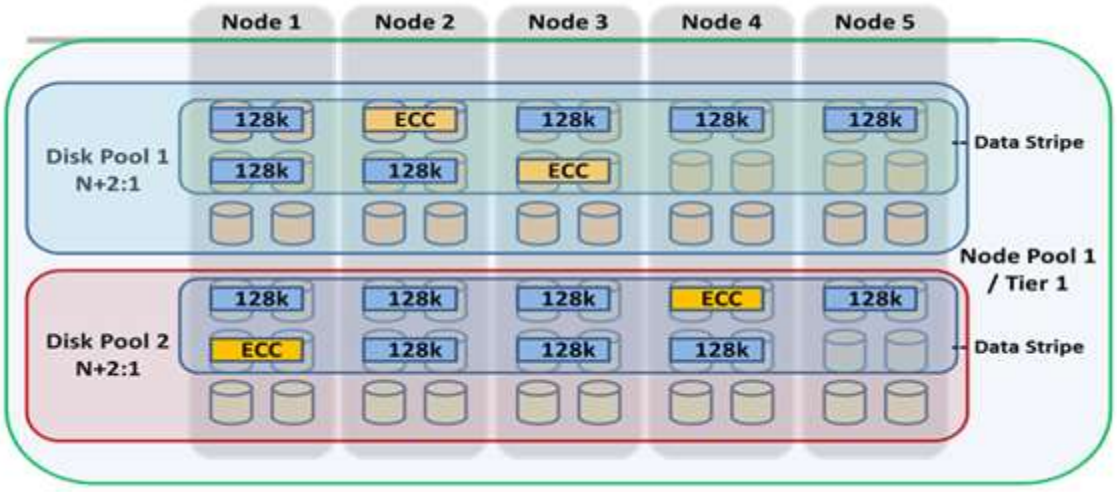

Figura 17: particionamento automático com o SmartPools

🕮 Mais informações estão disponíveis no white paper SmartPools.

As plataformas de hardware modular PowerScale de 6ª geração apresentam um projeto modular altamente denso, no qual quatro nós estão contidos em um único chassi de 4RU. Essa abordagem aprimora o conceito de pools de discos, pools de nós e “vizinhanças”, que adiciona outro nível de resiliência ao conceito de domínio de falha do OneFS. Cada chassi da plataforma de 6ª geração contém quatro módulos de computação (um por nó) e cinco contêineres de unidade, ou sleds, por nó.

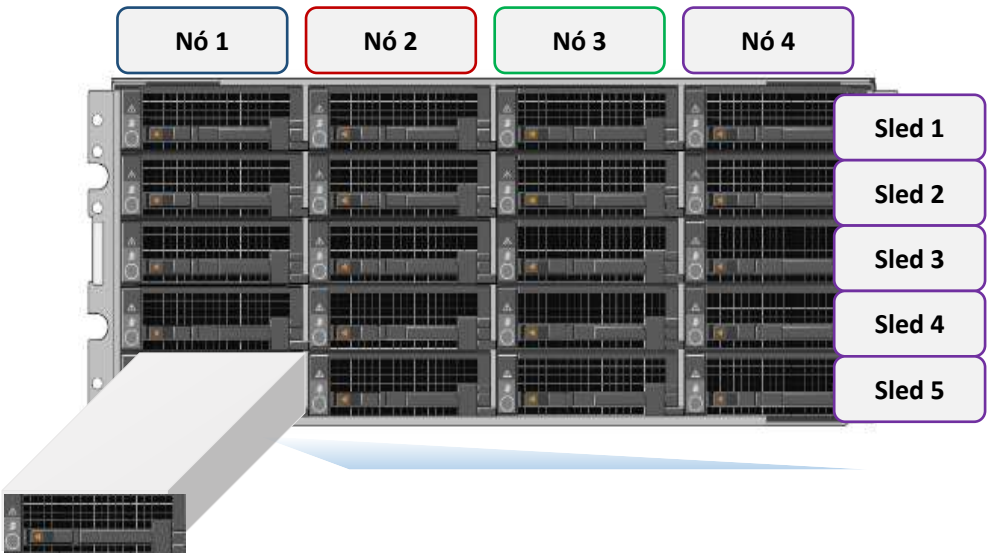

Figura 18. Visão frontal do chassi da plataforma de 6ª geração mostrando sleds da unidade.

Cada sled é uma bandeja que desliza na parte frontal do chassi e contém entre três e seis unidades, dependendo da configuração de um chassi específico. Os pools de discos são a menor unidade dentro da hierarquia de pools de armazenamento. O provisionamento do OneFS funciona no local na divisão de unidades dos nós semelhantes em conjuntos, ou pools de discos, com cada pool representando um domínio de falha separado. Esses pools de discos são protegidos por padrão em +2d:1n (ou a capacidade de resistir a duas unidades ou a uma falha de nó inteiro).

Os pools de discos são dispostos em todos os cinco sleds em cada nó de 6ª geração. Por exemplo, um nó com três unidades por sled terá a seguinte configuração de pool de discos:

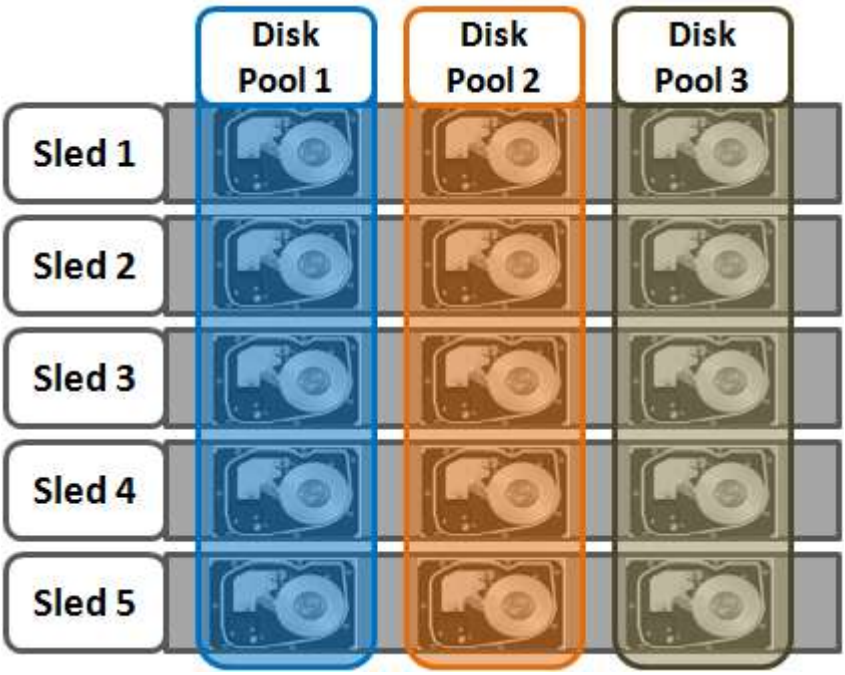

Figura 19. Pools de discos do OneFS

Os pools de nós são grupos de pools de discos, distribuídos entre os nós semelhantes de armazenamento (classes de compatibilidade). Isso é mostrado na Figura 20 abaixo. Vários grupos de diferentes tipos de nó podem trabalhar juntos em um cluster único e heterogêneo. Por exemplo: um pool de nós da série F para aplicativos com uso intenso de IOPS, um pool de nós da série H, principalmente usado para cargas de trabalho sequenciais e de alta simultaneidade, e um pool de nós da série A, principalmente usado para arquivamento near-line e/ou cargas de trabalho de arquivo morto.

Isso permite que o OneFS apresente um pool de recursos de armazenamento único composto por vários tipos de mídia de unidade – SSD, SAS de alta velocidade, SATA de grande capacidade etc. – oferecendo uma variedade de características diferentes de desempenho, proteção e capacidade. Esse pool de armazenamento heterogêneo, por sua vez, pode dar suporte a uma gama diversificada de aplicativos e requisitos de carga de trabalho com um só ponto de gerenciamento unificado. Ele também facilita a combinação de hardware mais antigo e mais recente, permitindo proteção de investimento simples, mesmo entre gerações de produtos, e atualizações contínuas de hardware.

Cada pool de nós contém apenas pools de discos do mesmo tipo de nós de armazenamento, e um pool de discos pode pertencer exatamente a um pool de nós. Por exemplo, nós da série F com unidades SSD de 1,6 TB estariam em um pool de nós, enquanto nós da série A com unidades SATA de 10 TB estariam em outro. Atualmente, é necessário um mínimo de 4 nós (um chassi) por pool de nós para hardware de 6ª geração, como o PowerScale H700, ou três nós por pool para nós autônomos, como o PowerScale F900.

"Vizinhanças" do OneFS são domínios de falha dentro de um pool de nós, e seu objetivo é aprimorar a confiabilidade em geral, além de proteger contra a indisponibilidade de dados a partir da remoção acidental dos sleds da unidade. Para nós autocontidos, como o PowerScale F200, o OneFS tem um tamanho ideal de 20 nós por pool de nós e um tamanho máximo de 39 nós. Com a adição do 40º nó, os nós são divididos em 2 vizinhanças de 20 nós.

Com a plataforma de 6ª geração, o tamanho ideal de um ambiente muda de 20 para 10 nós. Isso oferece proteção contra falhas simultâneas de registro do par de nós e falhas de chassi completo.

Nós parceiros são nós cujos registros são espelhados. Com a plataforma Gen6, em vez de cada nó armazenar seu registro em NVRAM como nas plataformas anteriores, os registros dos nós são armazenados nos SSDs, e cada registro tem uma cópia espelhada em outro nó. O nó que contém o registro espelhado é conhecido como nó parceiro. Há vários benefícios de confiabilidade obtidos com as alterações no registro. Por exemplo, os SSDs são mais persistentes e confiáveis do que NVRAM, o que requer uma bateria carregada para reter o estado. Além disso, com o registro espelhado, ambas as unidades de registro devem morrer antes de um registro ser considerado perdido. Dessa forma, a menos que as unidades de registro espelhado falhem, ambos os nós parceiros podem funcionar como de costume.

Com a proteção de nós parceiros, sempre que possível, os nós serão colocados em diferentes vizinhanças — e, portanto, domínios de falha diferentes. A proteção de nós parceiros é possível, uma vez que o cluster atinge 5 chassis completos (20 nós) quando, após a primeira divisão da vizinhança, o OneFS coloca nós parceiros em diferentes vizinhanças:

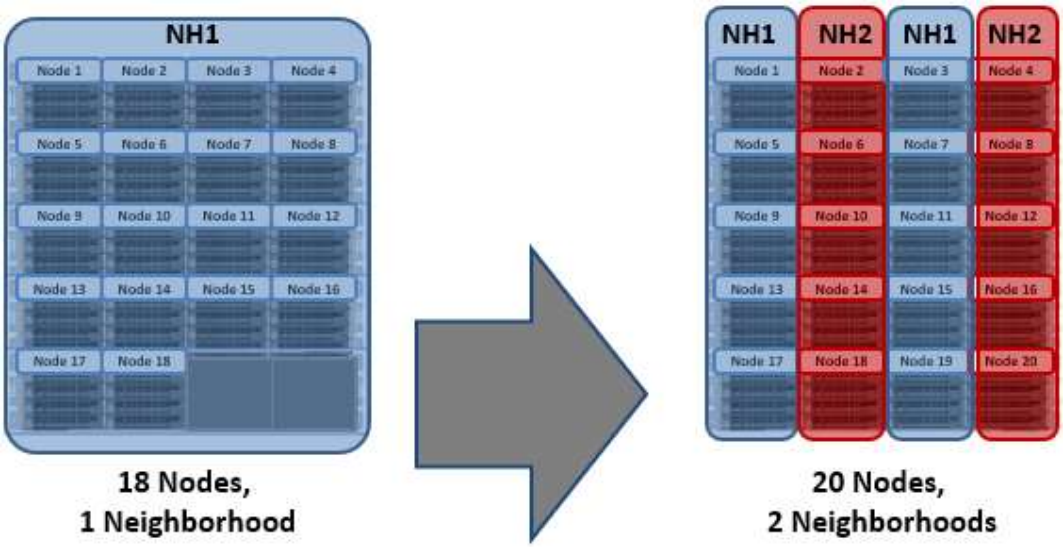

Figura 20. Divisão para 2 vizinhanças em 20 nós.

A proteção de nós parceiros aumenta a confiabilidade porque, se ambos os nós ficam inativos, eles estão em domínios de falha diferentes, portanto, os domínios de falha sofrem apenas a perda de um único nó.

Com a proteção do chassi, quando possível, cada um dos quatro nós de um chassi será colocado em um ambiente separado. A proteção do chassi torna-se possível em 40 nós, pois a divisão de vizinhança em 40 nós permite que todos os nós de um chassi sejam colocados em um ambiente diferente. Dessa forma, quando um cluster Gen6 de 38 nós é expandido para 40 nós, as duas vizinhanças existentes são divididas em 4 vizinhanças de 10 nós:

A proteção do chassi garante que, se um chassi inteiro falhar, cada domínio de falha só perderá um nó.

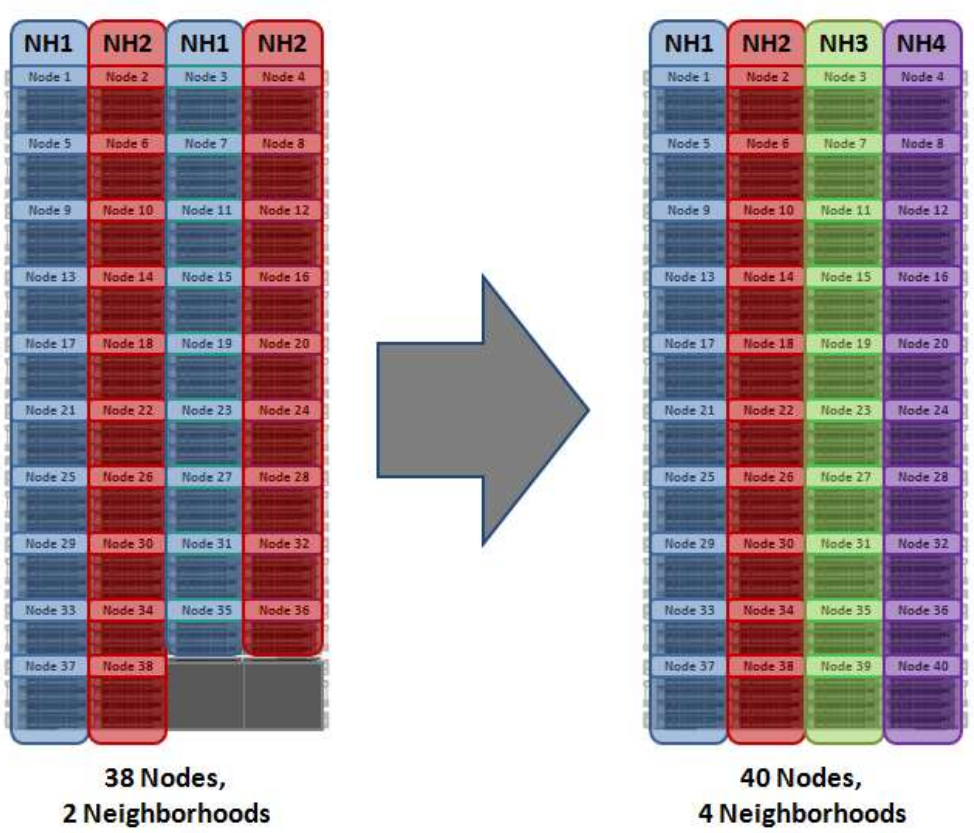

Figura 21. Vizinhanças do OneFS – quatro vizinhanças divididas.

ⓘ Um cluster de 40 nós ou maior com quatro vizinhanças, protegidas no nível padrão de +2d:1n pode sustentar uma falha de único nó por vizinhança. Isso protege o cluster contra uma falha de chassi único Gen6.

Em geral, um cluster da plataforma Gen6 terá uma confiabilidade de, pelo menos, uma ordem de magnitude maior do que os clusters da geração anterior de uma capacidade semelhante, como resultado direto dos seguintes aprimoramentos:

- Registros espelhados
- Vizinhanças menores
- Unidades de inicialização espelhada

Compatibilidade

Certos tipos de nós semelhantes, mas não idênticos, podem ser provisionados para um pool de nós existente por compatibilidade de nó. O OneFS exige que um pool de nós contenha no mínimo três nós.

ⓘ Devido a diferenças significativas de arquitetura, não há compatibilidades de nós entre a plataforma de 6ª geração, gerações anteriores de hardware ou os nós do PowerScale.

O OneFS também contém uma opção de compatibilidade com SSD, que permite que os nós com SSDs de capacidade diferente sejam provisionados para um só pool de nós.

A compatibilidade com SSD é criada e descrita na lista de compatibilidades com a WebUI do OneFS SmartPools e também é exibida na lista Níveis e pools de nós.

ⓘ Ao criar essa compatibilidade com SSD, o OneFS verifica automaticamente se os dois pools a serem mesclados têm o mesmo número de SSDs, nível, proteção solicitada e configurações de cache L3. Se essas configurações forem diferentes, a WebUI do OneFS solicitará a consolidação e o alinhamento dessas configurações.

🕮 Mais informações estão disponíveis no white paper SmartPools.

Protocolos compatíveis

Clients com credenciais e privilégios adequados podem criar, modificar e ler dados usando um dos métodos padrão aceitos para se comunicarem com o cluster:

- NFS (Network File System, sistema de arquivos de rede)
- SMB/CIFS (Server Message Block/Common Internet file system)
- FTP (File Transfer Protocol)
- HTTP (Hypertext Transfer Protocol)
- HDFS (Hadoop Distributed file system)
- API REST (Representational State Transfer Application Programming Interface)
- S3 (API de armazenamento em objeto)

Para o protocolo NFS, o OneFS é compatível com NFSv3 e NFSv4, além de NFSv4.1 no OneFS 9.3. Além disso, o OneFS 9.2 e versões posteriores incluem suporte para NFSv3overRDMA.

No Microsoft Windows, o protocolo SMB é compatível até a versão 3. Como parte do dialeto SMB3, o OneFS é compatível com os seguintes recursos:

- Múltiplos caminhos do SMB3
- Disponibilidade contínua e testemunha do SMB3
- Criptografia do SMB3

A criptografia do SMB3 pode ser configurada por compartilhamento, zona ou em todo o cluster. Apenas os sistemas operacionais compatíveis com a criptografia do SMB3 podem funcionar com compartilhamentos criptografados. Esses sistemas operacionais também podem funcionar com compartilhamentos não criptografados se o cluster estiver configurado para permitir conexões não criptografadas. Outros sistemas operacionais podem acessar compartilhamentos não criptografados apenas se o cluster estiver configurado para permitir conexões não criptografadas.

O root do file system de todos os dados no cluster é /ifs (o file system do OneFS). Isso pode ser apresentado por meio do protocolo SMB como um compartilhamento “ifs” (\\<cluster_name\ifs) e por meio do protocolo NFS como uma exportação “/ifs” (<cluster_name>:/ifs).

ⓘ Os dados são comuns entre todos os protocolos e, portanto, as alterações feitas no conteúdo do arquivo por meio de um protocolo de acesso são imediatamente visíveis a partir de todos os outros.

O OneFS oferece suporte completo a ambientes IPv4 e IPv6 em todas as redes Ethernet front-end, SmartConnect e o conjunto completo de protocolos de armazenamento e ferramentas de gerenciamento.

Além disso, o OneFS CloudPools é compatível com as APIs de armazenamento dos seguintes provedores de nuvem, permitindo que os arquivos sejam colocados em stub para uma série de destinos de armazenamento, inclusive:

- Amazon Web Services S3
- Microsoft Azure
- Serviço Google Cloud
- Alibaba Cloud
- Dell EMC ECS
- OneFS RAN (RESTful Access to Namespace)

📖 Mais informações estão disponíveis no guia de administração do CloudPools**.**

Operações não disruptivas – suporte a protocolos

O OneFS contribui para a disponibilidade dos dados dando suporte a failover e failback dinâmicos de NFSv3 w NFSv4 para clients Linux e UNIX, bem como disponibilidade contínua do SMB3 para clients Windows. Isso garante que, quando ocorre falha em um nó, ou quando é realizada uma manutenção preventiva, todas as gravações e leituras em trânsito são passadas para outro nó no cluster para concluir sua operação sem qualquer interrupção para o aplicativo ou usuário.

Durante um failover, os clients são redistribuídos igualmente entre todos os nós restantes no cluster para garantir o menor impacto possível sobre o desempenho. Se um nó é desativado por qualquer motivo, incluindo uma falha, os endereços IP virtuais nesse nó são perfeitamente migrados para outro nó no cluster.

Quando o nó off-line é colocado novamente on-line, o SmartConnect reequilibra automaticamente os clients NFS e SMB3 por todo o cluster para garantir máxima utilização de desempenho e armazenamento. Para manutenções periódicas do sistema e atualizações de software, essa funcionalidade permite rolling upgrades por nó que oferecem alta disponibilidade por toda a duração da janela de manutenção.

Filtragem de arquivos

A filtragem de arquivos do OneFS pode ser usada em vários clients NFS e SMB para permitir ou proibir gravações em uma zona de acesso, exportação ou de compartilhamento. Esse recurso impede certos tipos de extensões de arquivos que podem causar problemas de segurança, interrupções da produtividade, problemas de throughput ou desorganização do armazenamento.
A configuração pode ser feita por meio de uma lista de exclusões, que bloqueia extensões explícitas de arquivos, ou uma lista de inclusões, que permite explicitamente gravações apenas de determinados tipos de arquivos.

Desduplicação de dados – SmartDedupe

O produto SmartDedupe maximiza a eficiência de armazenamento de um cluster diminuindo a quantidade de armazenamento físico necessária para armazenar os dados de uma organização. A eficiência é alcançada verificando-se a existência de blocks idênticos nos dados em disco e, em seguida, eliminando-se as duplicações. Essa abordagem geralmente é chamada de desduplicação pós-processo ou assíncrona.

Depois que os blocks duplicados são detectados, o SmartDedupe move uma cópia única desses blocks para um conjunto especial de arquivos conhecidos como armazenamentos de sombra. Durante esse processo, os blocks duplicados são removidos dos arquivos reais e substituídos por ponteiros para os armazenamentos de sombra.

Com a desduplicação pós-processamento, os novos dados primeiro são armazenados no dispositivo de armazenamento e, em seguida, um processo subsequente analisa os dados em busca de semelhança. Isso significa que o desempenho de gravação ou modificação inicial do arquivo não é afetado, já que não é necessário nenhum cálculo adicional no caminho de gravação.

Arquitetura do SmartDedupe

A arquitetura do OneFS SmartDedupe é composta por cinco módulos principais:

- Caminho de controle de desduplicação

- Trabalhos de desduplicação
- Lógica de desduplicação
- Armazenamento de sombra
- Infraestrutura de desduplicação

O caminho de controle do SmartDedupe é composto pela WebUI (Web Management Interface, interface de gerenciamento da Web) do OneFS, CLI (Command Line Interface, interface de linha de comando) e API (Application Programming Interface, interface de programação de aplicativos) da plataforma RESTful, e é responsável por gerenciar a configuração, o agendamento e o controle do trabalho de desduplicação. O trabalho em si é um processo de segundo plano altamente distribuído que gerencia a orquestração da desduplicação em todos os nós do cluster. O controle de trabalho abrange a verificação de file system, a detecção e o compartilhamento de blocos de dados correspondentes, em sintonia com o mecanismo de desduplicação. A camada de infraestrutura de desduplicação é o módulo de kernel que executa a consolidação de blocos de dados compartilhados em armazenamentos de sombra, os contêineres de file system que mantêm tanto os blocos de dados físicos quanto as referências, ou indicadores, para blocks compartilhados. Esses elementos são descritos em mais detalhes abaixo.

Figura 22: arquitetura modular do OneFS SmartDedupe

📖 Mais informações estão disponíveis no white paper OneFS SmartDedupe.

Armazenamento de sombra

Os armazenamentos de sombra do OneFS são contêineres de file system que permitem que os dados sejam armazenados de maneira compartilhável. Dessa forma, os arquivos no OneFS podem conter tanto dados físicos quanto indicadores, ou referências, para blocks compartilhados em armazenamentos de sombra.

Os armazenamentos de sombra são semelhantes aos arquivos regulares, mas normalmente não contêm todos os metadados normalmente associados a inodes de arquivos regulares. Em particular, os atributos baseados em tempo (horário de criação, horário de modificação etc.) não são mantidos explicitamente. Cada armazenamento de sombra pode conter até 256 blocks, sendo que cada um pode ser referenciado por 32.000 arquivos. Se esse limite de referência de 32.000 for excedido, um novo armazenamento de sombra será criado. Além disso, os armazenamentos de sombra não fazem referência a outros armazenamentos de sombra. Além disso, os snapshots de armazenamentos de sombra não são permitidos, já que os armazenamentos de sombra não têm vínculos físicos.

ⓘ Os armazenamentos invisíveis também são utilizados para clones de arquivos do OneFS e SFSE (Small File Storage Efficiency, Eficiência de Armazenamento em Arquivo Pequeno), além da desduplicação.

Eficiência de armazenamento de pequenos arquivos

Outro consumidor principal de armazenamento de sombra é o recurso de eficiência de armazenamento de pequenos arquivos. Esse recurso maximiza a utilização de espaço de um cluster, diminuindo a quantidade de armazenamento físico necessária para armazenar os arquivos reduzidos que, muitas vezes, compõem um conjunto de dados de arquivamento, como encontrado nos fluxos de trabalho de PACS (Picture Archiving and Communication System, arquivamento de imagens e sistema de comunicação) da área de saúde.

A eficiência é alcançada examinando os dados em disco de arquivos reduzidos, que são protegidos por mirrors de cópia completa e os empacotando em armazenamentos de sombra. Esses armazenamentos de sombra são protegidos por paridade, em vez de espelhados, e normalmente fornecem eficiência de armazenamento de 80% ou mais.

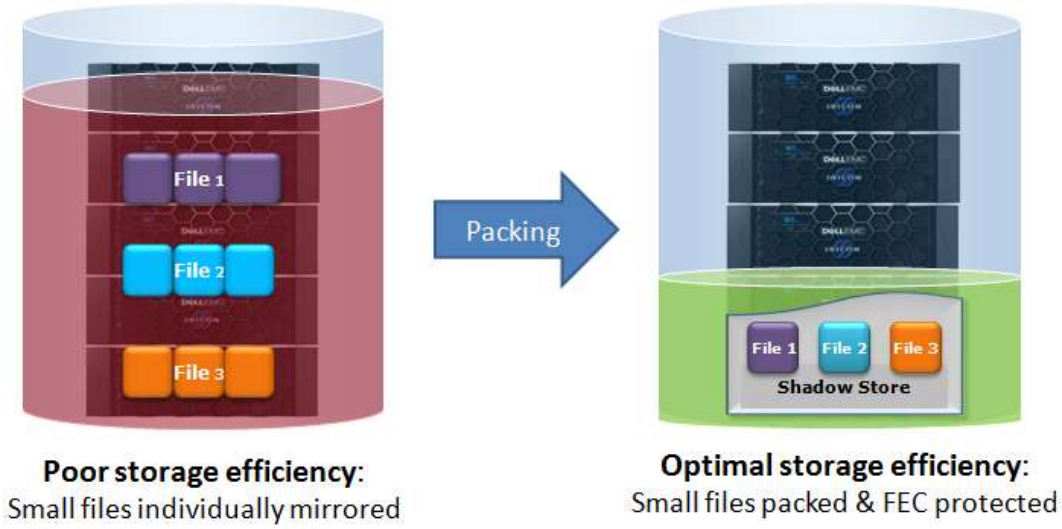

Figura 23: conteinerização de arquivos pequenos

O recurso de eficiência de armazenamento de pequenos arquivos troca uma pequena degradação no desempenho de latência de leitura por maior utilização do armazenamento. Os arquivos compactados obviamente permanecem graváveis, mas quando os arquivos em contêineres com referências de sombra são excluídos, truncados ou sobregravados, podem deixar blocks não referenciados em armazenamentos de sombra. Esses blocks são liberados posteriormente e podem resultar em espaços que reduzem a eficiência de armazenamento.

A perda de eficiência real depende do layout do nível de proteção usado pelo armazenamento de sombra. Os tamanhos menores dos grupos de proteção são mais suscetíveis, assim como os arquivos em contêineres, já que todos os blocks em contêineres têm no máximo um arquivo de referência e os tamanhos empacotados (tamanho do arquivo) são pequenos.

Um desfragmentador é fornecido para reduzir a fragmentação dos arquivos como resultado de sobregravações e exclusões. Esse desfragmentador de armazenamento de sombra é integrado ao trabalho ShadowStoreDelete. O processo de desfragmentação funciona dividindo cada arquivo em contêiner em fragmentos lógicos (aproximadamente 32 MB cada) e avaliando cada fragmento para verificar se há fragmentação.

Se a eficiência de armazenamento de um fragmento fragmentado está abaixo do destino, esse fragmento é processado evacuando os dados para outro local. A meta padrão de eficiência é 90% da eficiência máxima de armazenamento disponível com o nível de proteção usado pela área de armazenamento de sombra. Tamanhos maiores de grupos de proteção podem tolerar um nível mais alto de fragmentação antes que a eficiência de armazenamento fique abaixo desse limite.

Redução de dados em linha

A redução de dados em linha do OneFS está disponível nos nós All-Flash F900, F810, F600 e F200, nos chassis híbridos H700/7000 e H5600, bem como na plataforma de arquivamento A300/3000. A arquitetura do OneFS é composta pelos seguintes componentes principais:

- Plataforma de redução de dados
- Mecanismo de compactação e mapeamento de fragmentos

- Fase de remoção de blocks com valor zero
- Índice de desduplicação em memória e infraestrutura de área de armazenamento de sombra
- Estrutura de geração de relatórios e alertas de redução de dados
- Caminho de controle da redução de dados

O caminho de gravação da redução de dados em linha é composto por três fases principais:

- Remoção de blocks com valor zero
- Desduplicação em linha
- Compactação em linha

Se a compactação e a desduplicação em linha estiverem ativadas em um cluster, a remoção de blocks com valor será executada primeiramente, seguida pela desduplicação e, em seguida, pela compactação. Esse pedido permite que cada fase reduza o escopo do trabalho em cada fase subsequente.

Figura 24: fluxo de trabalho da redução de dados em linha

O F810 inclui um recurso de liberação de compactação de hardware, com cada nó em um chassi do F810 contendo um adaptador Flex Mellanox Innova-2. Isso significa que a compactação e a descompactação são realizadas de maneira transparente pelo adaptador Mellanox com latência mínima, evitando, assim, a necessidade de consumir os caros recursos de CPU e memória de um nó.

O mecanismo de compactação de hardware do OneFS usa zlib, com uma implementação de software do igzip para os nós PowerScale F900, F810, F600, F200, H700/7000, H5600 e os nós A300/3000. A compactação de software também é usada como restauração em caso de falha de hardware de compactação e em um cluster misto para uso em nós que não são do F810 sem um recurso de compactação de hardware e como restauração em caso de falha de hardware de compactação. O OneFS emprega um tamanho de fragmento de compactação de 128 KB, sendo que cada fragmento consiste em dezesseis blocos de dados de 8 KB. Isso é ideal, já que também é o mesmo tamanho que o OneFS usa para suas unidades de fração de proteção de dados, proporcionando simplicidade e eficiência, evitando a sobrecarga do pacote do fragmento adicional.

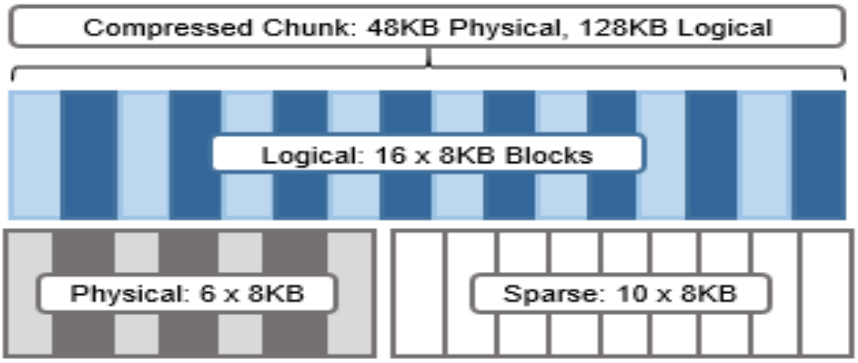

Figura 25: partes de compactação e sobreposição transparente do OneFS

Considere o diagrama acima. Depois da compactação, esse fragmento é reduzido de 16 para 6 blocks de 8 KB de tamanho. Isso significa que o fragmento agora está fisicamente com um tamanho de 48 KB. O OneFS fornece uma sobreposição lógica transparente para os atributos físicos. Essa sobreposição descreve se os dados de backup são compactados ou não e quais blocks no fragmento são físicos ou fragmentados, de modo que os consumidores do file system não são afetados pela compactação. Dessa forma, o fragmento compactado é logicamente representado como 128 KB de tamanho, independentemente de seu tamanho físico real.

A economia de eficiência deve ser de pelo menos 8 KB (um block) para que a compactação ocorra, caso contrário, o fragmento ou o arquivo será transmitido e permanecerá no estado original descompactado. Por exemplo, um arquivo de 16 KB que produz 8 KB (um block) de economia seria compactado. Depois que um arquivo tiver sido compactado, ele ficará protegido por FEC.

Fragmentos de compactação nunca atravessarão pools de nós. Isso evita a necessidade de descompactar ou recompactar dados para alterar os níveis de proteção, realizar gravações recuperadas ou, de outro modo, transferir os limites do grupo de proteção.

Dimensionamento dinâmico/dimensionamento sob demanda

Desempenho e capacidade

Ao contrário dos sistemas de armazenamento independentes que devem fazer “scale-out” quando é preciso obter mais desempenho ou capacidade, o OneFS permite que um sistema de armazenamento seja “dimensionado horizontalmente”, aumentando perfeitamente o volume ou o file system existente em petabytes de capacidade ao mesmo tempo que aumenta o desempenho em conjunto de maneira linear.

Adicionar recursos de desempenho e capacidade a um cluster é significativamente mais fácil do que com outros sistemas de armazenamento, exigindo apenas três passos simples do administrador de armazenamento: adicionar outro nó ao rack, conectar o nó à rede de back-end e instruir o cluster a acrescentar o nó adicional. O novo nó fornece capacidade e desempenho adicionais, pois cada nó inclui CPU, memória, cache, rede, NVRAM e caminhos de controle de E/S.

O recurso AutoBalance do OneFS moverá dados automaticamente através da rede de back-end de forma automática e coerente para que os dados existentes que residem no cluster se movam para este novo nó de armazenamento. Esse rebalanceamento automático garante que o novo nó não vá se tornar um ponto de acesso para novos dados e que os dados existentes sejam capazes de obter os benefícios de um sistema de armazenamento mais avançado. O recurso AutoBalance do OneFS também é completamente transparente para o usuário final e pode ser ajustado para minimizar o impacto nas cargas de trabalho de alto desempenho. Esse recurso por si só permite que o OneFS seja dimensionado de modo transparente e dinâmico, de TBs para mais de PBs sem tempo extra de gerenciamento para o administrador nem maior complexidade no sistema de armazenamento.

Um sistema de armazenamento em grande escala deve oferecer o desempenho necessário para diversos fluxos de trabalho, sejam eles sequenciais, simultâneos ou aleatórios. Haverá diferentes fluxos de trabalho entre aplicativos e em aplicativos individuais. O OneFS atende a todas essas necessidades simultaneamente com um software inteligente. Mais importante ainda, com o OneFS o throughput e a IOPS são dimensionados linearmente com o número de nós presentes em um sistema único. Devido à distribuição de dados equilibrada, ao rebalanceamento automático e ao processamento distribuído, o OneFS é capaz de utilizar CPUs adicionais, portas de rede e memória à medida que o sistema aumenta.

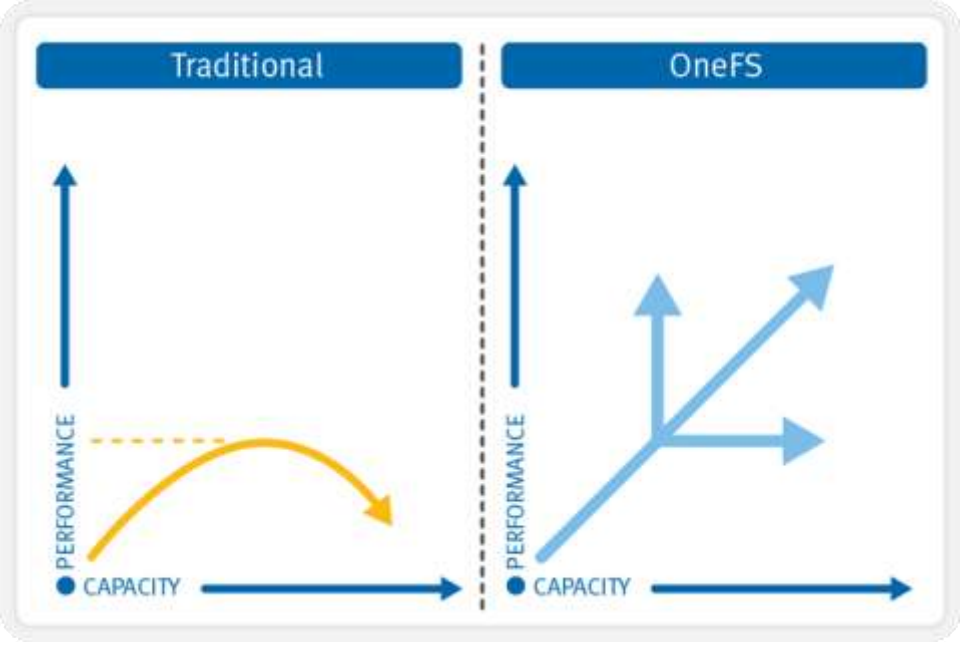

Figura 26: escalabilidade linear do OneFS

Interfaces

Os administradores podem usar várias interfaces para administrar um cluster de armazenamento em seus ambientes:

- Interface de usuário de administração da Web (“WebUI”)
- CLI (Command Line Interface, interface de linha de comando) via acesso à rede do SSH ou conexão serial RS232
- Painel LCD em nós próprios para funções simples de adição/remoção
- API de plataforma RESTful para controle e automatização programáticos de configuração e gerenciamento de cluster.

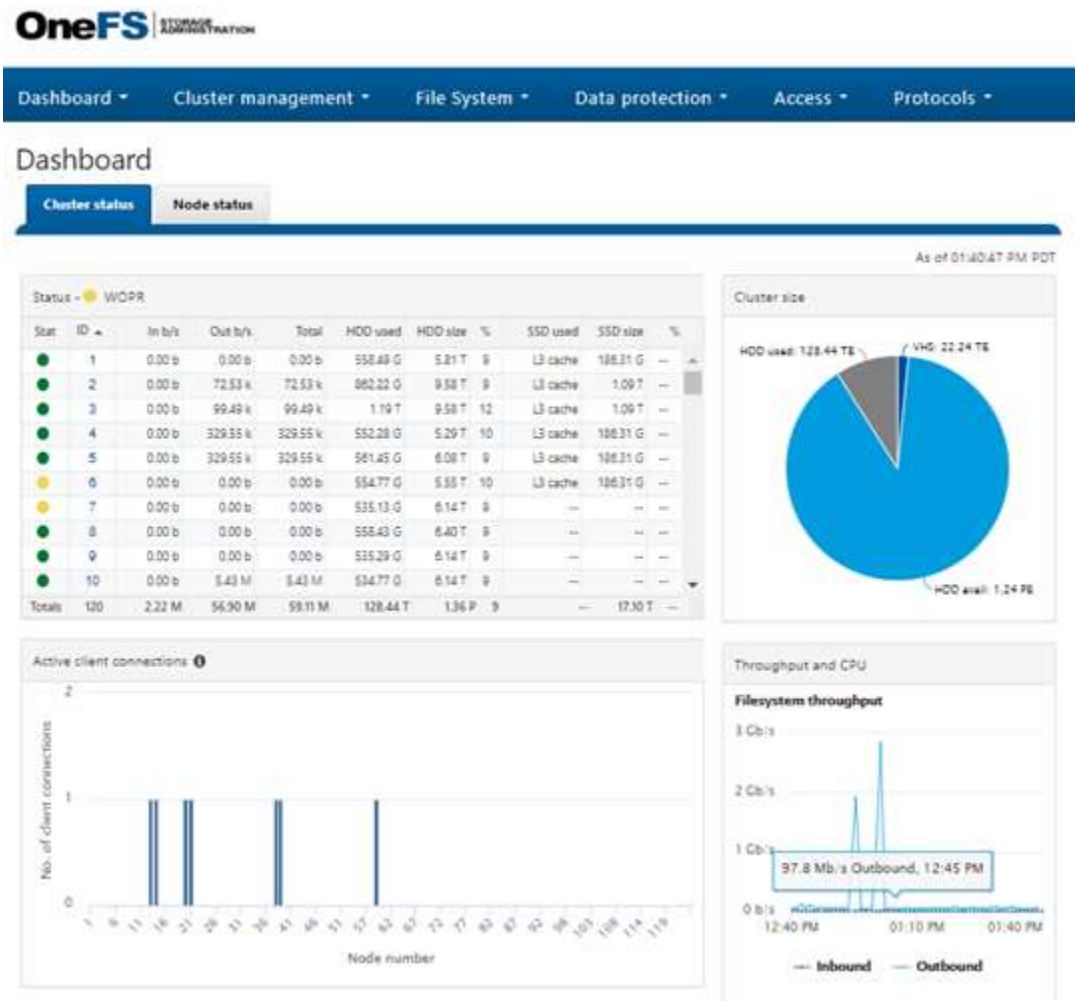

Figura 27: interface de usuário da Web do OneFS

📖 Mais informações sobre a configuração de comandos e recursos do OneFS estão disponíveis no Guia de administração do OneFS.

Autenticação e controle de acesso

Os serviços de autenticação oferecem uma camada de segurança ao verificar as credenciais dos usuários antes de permitir que eles acessem e modifiquem arquivos. O OneFS dá suporte a quatro métodos para autenticar usuários:

- Active Directory (AD)
- LDAP (Lightweight Directory Access Protocol)
- NIS (Network Information Service)
- Local users & Groups

O OneFS dá suporte ao uso de mais de um tipo de autenticação. No entanto, é recomendado que você compreenda plenamente as interações entre os tipos de autenticação antes de permitir múltiplos métodos no cluster. Consulte a documentação do produto para obter informações detalhadas sobre como configurar corretamente os modos de autenticação.

Active Directory

O Active Directory, uma implementação de LDAP da Microsoft, é um serviço de diretório que pode armazenar informações sobre os recursos de rede. Enquanto o Active Directory pode atender a muitas funções, a principal razão para a adesão do cluster ao domínio é realizar a autenticação de usuários e grupos.

Você pode definir e gerenciar as configurações de um cluster do Active Directory a partir da interface de administração da Web ou da interface de linha de comando, no entanto, é recomendado que você use a Administração da Web sempre que possível.

Cada nó no cluster compartilha a mesma conta de máquina do Active Directory, o que o torna muito fácil de administrar e gerenciar.

LDAP

O LDAP (Lightweight Directory Access Protocol) é um protocolo de rede utilizado para a definição, consulta e modificação de serviços e recursos. A principal vantagem do LDAP é a natureza aberta dos serviços de diretório e a capacidade de usar LDAP em várias plataformas. O sistema de armazenamento colocado em ambiente de cluster pode usar LDAP para autenticação de usuários e grupos a fim de conceder-lhes acesso ao cluster.

NIS

O NIS (Network Information Service, serviço de informação da rede), projetado pela Sun Microsystems, é um protocolo de serviços de diretório que pode ser usado pelo OneFS para autenticar usuários e grupos ao acessar o cluster. O NIS, por vezes chamado de YP (Yellow Pages), é diferente do NIS+, que o cluster do OneFS não aceita.

Usuários locais

O OneFS oferece suporte à autenticação de usuário e grupo local. Você pode criar contas de usuário local e grupo diretamente no cluster, utilizando a interface WebUI. A autenticação local pode ser útil quando os serviços de diretório – Active Directory, LDAP ou NIS – não são usados ou quando um usuário específico ou aplicativo precisa acessar o cluster.

Zonas de acesso

As zonas de acesso fornecem um método para particionar logicamente o acesso ao cluster e alocar recursos para unidades independentes, proporcionando, assim, um ambiente compartilhado de tenant ou vários tenants. Para facilitar isso, as zonas de acesso unem os três componentes principais de acesso externo:

- Configuração de rede do cluster
- Acesso a protocolo de arquivo
- Autenticação

Desse modo, as zonas do SmartConnect estão associadas a um conjunto de compartilhamentos SMB, exportações NFS, racks HDFS e um ou mais provedores de autenticação para fins de controle de acesso. Isso oferece os benefícios de um file system único gerenciado centralmente, que pode ser provisionado e protegido para vários tenants. Isso é particularmente útil para ambientes corporativos, onde várias unidades de negócios separadas são atendidas por um departamento central de TI. Outro exemplo é durante uma iniciativa de consolidação de servidores, ao mesclar vários servidores de arquivos do Windows que fazem parte de florestas separadas e não confiáveis do Active Directory.

Com zonas de acesso, a zona de acesso incorporada ao sistema inclui uma instância de cada provedor de autenticação aceito, todos os compartilhamentos SMB disponíveis e todas as exportações NFS disponíveis por padrão.

Esses provedores de autenticação podem incluir várias instâncias do Microsoft Active Directory, LDAP, NIS e bancos de dados de usuário local ou de grupo.

Administração com base em funções

A administração baseada em funções é um sistema de RBAC (Roles Based Access Control, Controle de Acesso Baseado em Função) de gerenciamento de cluster que divide os poderes dos usuários “raiz” e “administrador” em mais privilégios granulares e permite sua atribuição a funções específicas. Essas funções podem ser concedidas a outros usuários não privilegiados. Por exemplo, a equipe de operações de data center pode ter direitos de somente leitura em todo o cluster, permitindo o acesso de monitoramento completo, mas sem alterações de configuração a serem feitas. O OneFS fornece um conjunto de funções integradas, incluindo administrador de segurança e sistema de auditoria, além da capacidade de criar funções definidas personalizadas, por zona de acesso ou em todo o cluster. A administração baseada em funções é integrada à interface de linha de comando, à WebUI e à API de plataforma do OneFS.

📖 Para obter mais informações sobre gerenciamento de identidades, autenticação e controle de acesso em ambientes multiprotocolo, consulte o guia de segurança multiprotocolo do OneFS.

Auditoria do OneFS

O OneFS fornece a capacidade de auditar a configuração do sistema e a atividade dos protocolos NFS, SMB e HDFS em um cluster. Isso permite que as organizações atendam a várias exigências de governança de dados e conformidade regulamentar com as quais podem estar vinculadas.

Todos os dados de auditoria são armazenados e protegidos no file system do cluster e organizados por tópicos de auditoria. A partir daí, os dados de auditoria podem ser exportados por meio do framework Dell EMC CEE (Common Event Enabler) para aplicativos de terceiros, como o varonis DatAdvantage e o Symantec Data Insight. A auditoria de protocolo OneFS pode ser habilitada por zona de acesso, permitindo o controle granular em todo o cluster.

Um cluster pode gravar eventos de auditoria em até cinco servidores CEE por nó em uma configuração paralela com balanceamento de carga. Isso permite que o OneFS ofereça uma solução corporativa completa de auditoria.

📖 Mais informações estão disponíveis no white paper Auditoria do OneFS.

Upgrade de software

O upgrade para a versão mais recente do OneFS permite que você tire proveito de todos os novos recursos, correções e funcionalidades. Os clusters podem receber upgrade usando dois métodos: simultaneamente ou upgrade sem interrupção

Upgrade simultâneo

O upgrade simultâneo instala o novo sistema operacional e reinicializa todos os nós do cluster ao mesmo tempo. O upgrade simultâneo requer uma interrupção temporária, menos de dois minutos, do serviço durante o processo de upgrade enquanto os nós são reiniciados.

Rolling upgrade

Um rolling upgrade faz o upgrade e reinicia individualmente cada nó no cluster de modo sequencial. Durante um rolling upgrade, o cluster permanece on-line e continua servindo dados a clients sem interrupção no serviço. Em versões anteriores do OneFS 8.0, um rolling upgrade só pode ser realizado dentro de uma família de versão de código do OneFS e não entre revisões principais de versão de código do OneFS. A partir do OneFS 8.0, todas as novas versões terão rolling upgrades a partir da versão anterior.

Upgrades não disruptivos

NDUs (Non-Disruptive Upgrades, upgrades não disruptivos) permitem que um administrador de cluster atualize o sistema operacional de armazenamento enquanto seus usuários finais continuam acessando os dados sem erros nem interrupções. Atualizar o sistema operacional em um cluster é uma questão simples de um rolling upgrade. Durante esse processo, um nó por vez recebe upgrade para o novo código, e os clients ativos NFS e SMB3 conectados a ele são migrados automaticamente para outros nós no cluster. Também é permitido o upgrade parcial, no qual é possível fazer upgrade de um subconjunto de nós do cluster. O subconjunto de nós também pode ser aumentado durante o upgrade. Um upgrade pode ser pausado e retomado, permitindo que os clientes façam upgrades em várias janelas de manutenção menores. Além disso, o OneFS 8.2.2 e versões posteriores oferecem upgrades paralelos, nos quais os clusters podem fazer upgrade de uma vizinhança inteira ou um domínio de falha por vez, reduzindo substancialmente a duração dos upgrades de clusters grandes. O OneFS 9.2 e versões posteriores combinam upgrades de sistema operacional e firmware, reduzindo substancialmente o impacto e a duração dos upgrades, permitindo que eles ocorram em conjunto. O 9.2 e as versões posteriores também incluem upgrades baseados em drenagem, nos quais os nós são impedidos de reinicializar ou reiniciar os serviços de protocolo até que todos os clients SMB tenham se desconectado do nó.

Capacidade de reversão

O OneFSá suporte a reversão de upgrade, oferecendo a capacidade de retornar um cluster com um upgrade não confirmado à versão anterior do OneFS.

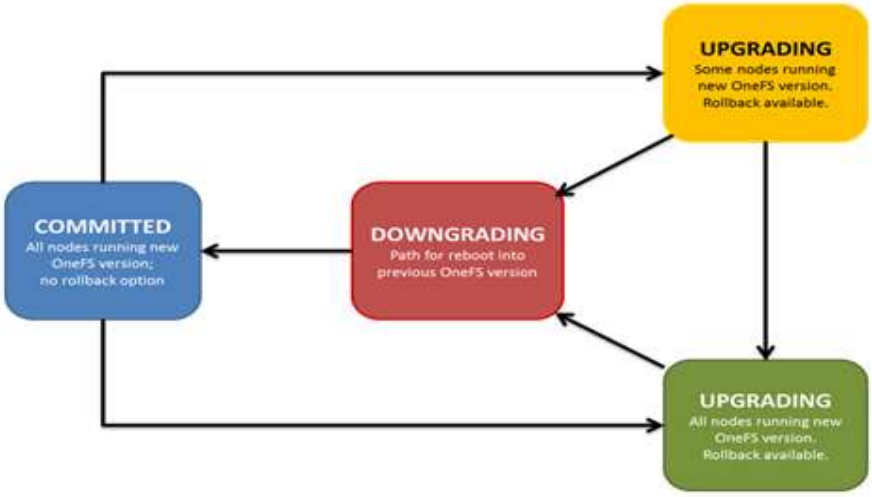

Figura 28: Estados de upgrade não disruptivo do OneFS

Atualizações automáticas de firmware

Os clusters do OneFS são compatíveis com atualizações automáticas de firmware de unidade para unidades novas e de substituição, como parte do processo de atualização de firmware sem interrupções. Atualizações de firmware são fornecidas por meio de pacotes de suporte a unidades, que simplificam e agilizam o gerenciamento de unidades existentes e novas em todo o cluster. Isso garante que o firmware da unidade esteja atualizado e reduz a probabilidade de falhas devido a problemas conhecidos de unidade. Dessa forma, as atualizações automáticas do firmware da unidade são um componente importante da estratégia de operações de alta disponibilidade e sem interrupções do OneFS. O firmware de unidade e de nó pode ser aplicado como um rolling upgrade ou por meio de uma reinicialização completa do cluster.

Antes do OneFS 8.2, as atualizações de firmware de nó precisavam ser instaladas em um nó de cada vez, o que era uma operação demorada, especialmente em clusters grandes. As atualizações de firmware de nó agora podem ser coreografadas em um cluster, fornecendo uma lista de nós que serão atualizados simultaneamente. A ferramenta auxiliar de upgrade pode ser usada para selecionar uma combinação desejada de nós que podem ser atualizados simultaneamente e uma lista explícita de nós que não devem ser atualizados juntos (por exemplo, nós em um par de nós).

Executando o upgrade

Como parte de um upgrade, o OneFS executa automaticamente uma verificação pré-instalação. Isso verifica se a configuração em sua instalação atual do OneFS é compatível com a versão do OneFS a que o upgrade se destina. Quando uma configuração não compatível é encontrada, o upgrade é interrompido e são exibidas instruções sobre a solução do problema. Executar proativamente a verificação do upgrade pré-instalação antes de iniciar um upgrade ajuda a evitar qualquer interrupção devido a uma configuração incompatível.

Software OneFS de gerenciamento e proteção de dados

O OneFS oferece um portfólio abrangente de software de proteção de dados e gerenciamento para atender a suas necessidades:

Módulos de software	Função	Descrição
CloudIQ™	**Monitoramento de integridade do cluster**	Implemente lógica analítica inteligente e preditiva para monitorar proativamente a integridade do cluster.
InsightIQ™	**Gerenciamento de desempenho**	Maximize o desempenho do cluster com ferramentas inovadoras de monitoramento de desempenho e geração de relatórios
DataIQ™	**Análise e gerenciamento de dados**	Localize, acesse e gerencie dados em segundos, não importando onde eles estejam, no armazenamento em file e armazenamento em object, no local ou na nuvem. Obtenha uma visão abrangente dos sistemas de armazenamento heterogêneo com apenas um painel, efetivamente dividindo os dados interrompidos em silos.
SmartPools™	**Gerenciamento de recursos**	Implemente uma estratégia de armazenamento hierárquico, automatizado e altamente eficiente para otimizar o desempenho e os custos de armazenamento
SmartQuotas™	**Gerenciamento de dados**	Atribua e gerencie quotas que perfeitamente particionam e fazem o provisionamento thin do armazenamento em segmentos fáceis de gerenciar nos níveis de cluster, diretório, subdiretório, usuário e grupo
SmartConnect™	**Acesso a dados**	Habilite o balanceamento de carga de conexão de clients, bem como o failover e o failback de NFS dinâmicos de conexões de clients em nós de armazenamento para otimizar o uso de recursos do cluster
SnapshotIQ™	**Proteção de dados**	Proteja dados com eficiência e confiabilidade utilizando snapshots seguros e quase instantâneos causando pouca ou nenhuma sobrecarga de desempenho. Acelere a recuperação de dados críticos com restaurações de snapshots quase imediatas sob demanda. Crie cópias modificáveis e com uso eficiente de espaço e tempo de um snapshot somente leitura com snapshots graváveis do OneFS.
SyncIQ™	**Replicação de dados**	Replique e distribua conjuntos de dados grandes e essenciais de modo assíncrono para vários sistemas de armazenamento compartilhados em vários locais para obter o recurso de recuperação de desastres confiável. Simplicidade de failover e failback controlados por botões para aumentar a disponibilidade de dados essenciais.
SmartLock™	**Retenção de dados**	Proteja seus dados críticos contra alterações ou exclusões acidentais, prematuras ou mal-intencionadas com nossa abordagem a WORM (Write Once Read Many times) baseada em software e atenda aos exigentes requisitos de conformidade e governança, como a SEC 17a-4.
SmartDedupe™	**Desduplicação de dados**	Maximize a eficiência do armazenamento analisando o cluster em busca de blocks idênticos e, em seguida, eliminando as duplicações, diminuindo a quantidade de armazenamento físico necessário.
CloudPools™	**Classificação em nuvem**	O CloudPools permite que você defina quais dados em seu cluster devem ser arquivados no armazenamento em nuvem. Os provedores de serviços em nuvem incluem Microsoft Azure, Google Cloud, Amazon S3, Dell EMC ECS e OneFS nativo.

Tabela 3: portfólio de serviços de dados do Dell EMC PowerScale

Consulte a documentação do produto para obter mais detalhes.

Conclusão

Com as soluções NAS de scale-out da Dell EMC habilitadas pelo sistema operacional OneFS, as organizações podem dimensionar de TBs para PBs em um só file system, um só volume, com um ponto único de administração. O OneFS oferece alto desempenho, alto throughput ou ambos, sem adicionar complexidade de gerenciamento.

Os data centers de última geração devem ser desenvolvidos para fins de dimensionamento sustentável. Eles vão aproveitar o poder da automatização e a comoditização do hardware, assegurar o consumo total do fabric de rede e proporcionar o máximo de flexibilidade para organizações que pretendem satisfazer a um conjunto de requisitos em constante mudança.

O OneFS é o file system de última geração projetado para atender a esses desafios. O OneFS fornece:

- Sistema de arquivos único totalmente distribuído
- Cluster com alto desempenho e totalmente simétrico
- Particionamento de arquivos em todos os nós em um cluster
- Software automatizado para eliminar a complexidade
- Balanceamento dinâmico de conteúdo
- Proteção flexível de dados
- Alta disponibilidade
- Administração de linha de comando e baseada na Web

O OneFS é ideal para aplicativos de “Big Data” não estruturados com base em arquivos em ambientes corporativos de data lake — inclusive diretórios base de grande escala, compartilhamentos de arquivos, arquivamentos, virtualização e lógica analítica de negócios —, bem como para uma grande variedade de ambientes de computação de alto desempenho com uso intenso de dados, inclusive exploração de energia, serviços financeiros, serviços de hospedagem e Internet, Business Intelligence, engenharia, manufatura, mídia e entretenimento, bioinformática e pesquisa científica.

DÊ O PRÓXIMO PASSO

Entre em contato com seu representante de vendas ou revendedor autorizado da Dell EMC para saber mais sobre como as soluções de armazenamento NAS do PowerScale podem beneficiar sua organização.

Acesse Dell EMC PowerScale para comparar recursos e obter mais informações.

Saiba mais sobre as soluções do Dell EMC PowerScale

Entre em contato com um especialista da Dell EMC

Veja mais recursos

Participe da conversa sobre #DellEMCStorage

Número de referência: H10719.14