LIVRE BLANC

# DELL EMC POWERSCALE ONEFS : PRÉSENTATION TECHNIQUE


## Résumé

Ce livre blanc fournit des informations techniques sur les principales caractéristiques et fonctionnalités du système d'exploitation OneFS qui est utilisé pour alimenter toutes les solutions de stockage NAS scale-out Dell EMC PowerScale.



Septembre 2021

# Révisions

| Version | Date | Commentaire |
|---|---|---|
| **1.0** | Novembre 2013 | Version originale de OneFS 7.1 |
| **2.0** | Juin 2014 | Mise à jour vers OneFS 7.1.1 |
| **3.0** | Novembre 2014 | Mise à jour vers OneFS 7.2 |
| **4.0** | Juin 2015 | Mise à jour vers OneFS 7.2.1 |
| **5.0** | Novembre 2015 | Mise à jour vers OneFS 8.0 |
| **6.0** | Septembre 2016 | Mise à jour vers OneFS 8.0.1 |
| **7.0** | Avril 2017 | Mise à jour pour OneFS 8.1 |
| **8.0** | Novembre 2017 | Mise à jour pour OneFS 8.1.1 |
| **9.0** | Février 2019 | Mise à jour vers OneFS 8.1.3 |
| **10.0** | Avril 2019 | Mise à jour pour OneFS 8.2 |
| **11.0** | Août 2019 | Mis à jour pour OneFS 8.2.1 |
| **12.0** | Décembre 2019 | Mise à jour pour OneFS 8.2.2 |
| **13.0** | Juin 2020 | Mise à jour pour OneFS 9.0 |
| **14.0** | Septembre 2020 | Mise à jour pour OneFS 9.1 |
| **15.0** | Avril 2021 | Mise à jour pour OneFS 9.2 |
| **16.0** | Septembre 2021 | Mise à jour pour OneFS 9.3 |

# Remerciements

Ce livre blanc a été conçu par les éléments suivants :

Auteur : Nick Trimbee

# SOMMAIRE

# Introduction

Les trois couches du modèle de stockage traditionnel (système de fichiers, gestionnaire de volume et protection des données) ont évolué au fil du temps pour s'adapter aux besoins des architectures de stockage à petite échelle, mais présentent plus de complexité et ne sont pas très adaptées aux systèmes capables d'évoluer vers plusieurs pétaoctets. Le système d'exploitation OneFS remplace tous ces éléments et fournit un système de fichiers en cluster unifié doté d'une protection des données évolutive intégrée, et permet de ne pas avoir besoin de gérer les volumes. OneFS est un bloc de construction fondamental pour les infrastructures scale-out, garantissant une importante évolutivité et une très grande efficacité. Il alimente toutes les solutions de stockage NAS Dell EMC PowerScale.

De par sa conception, OneFS s'adapte non seulement aux machines, mais également au personnel, permettant aux systèmes à grande échelle d'être gérés avec une fraction du personnel requis pour les systèmes de stockage traditionnels. OneFS élimine la complexité et intègre des fonctionnalités d'autoréparation et d'autogestion qui réduisent considérablement les tâches de gestion du stockage. OneFS intègre également le parallélisme à un niveau très profond du système d'exploitation, de sorte que presque tous les services système clés sont distribués entre plusieurs unités de matériel. Ainsi, OneFS peut évoluer dans pratiquement toutes les dimensions à mesure que l'infrastructure se développe, garantissant que les stratégies qui fonctionnent aujourd'hui continueront de fonctionner à mesure que la taille du Dataset augmente.

OneFS est un système de fichiers parfaitement symétrique sans point unique de défaillance qui tire parti du clustering pour adapter les performances et la capacité, mais aussi pour permettre un basculement sur incident de/vers n'importe quel système et plusieurs niveaux de redondance qui vont bien au-delà des capacités RAID. La tendance, pour les sous-systèmes de disques, est une augmentation lente des performances et une augmentation rapide des densités de stockage. OneFS répond à cette réalité en faisant évoluer la quantité de redondance et la vitesse de réparation des pannes. Cela permet à OneFS d'évoluer jusqu'à plusieurs pétaoctets tout en offrant plus de fiabilité que les petits systèmes de stockage traditionnels.

Le matériel PowerScale fournit l'appliance sur laquelle OneFS est exécuté. Les composants matériels sont innovants, mais universels, ce qui garantit les bénéfices des courbes d'efficacité et de coût en constante amélioration du matériel générique. OneFS permet au matériel d'être incorporé ou retiré du cluster à volonté et à tout moment, en « soustrayant » les données et les applications de ce matériel. Les données possèdent une longévité illimitée et sont protégées contre les vicissitudes liées à l'évolution du matériel. Le coût élevé et les difficultés liées aux migrations de données, de même que l'actualisation du matériel, sont éliminés.

OneFS est particulièrement adapté aux applications de « Big Data » basées sur des fichiers ou des données non structurées dans des environnements d'entreprise : répertoires de base à grande échelle, partages de fichiers, archives, virtualisation et analytique métier. De ce fait, OneFS est très répandu dans de nombreux secteurs gérant à l'heure actuelle d'importants volumes de données, comme l'énergie, les services financiers, Internet et les services d'hébergement, la business intelligence, l'ingénierie, la fabrication, les médias et le divertissement, la bio-informatique, la recherche scientifique et d'autres environnements informatiques haute performance.

# Audience visée

Ce livre blanc partage des informations sur le déploiement et la gestion d'un cluster Dell EMC PowerScale et fournit un descriptif complet de l'architecture OneFS.

Il s'adresse à quiconque configure et gère un environnement de stockage en cluster PowerScale. Il suppose que le lecteur dispose de connaissances de base sur le stockage, la gestion de réseau, les systèmes d'exploitation et la gestion des données.

📖 Pour plus d'informations sur la configuration des commandes et des fonctionnalités de OneFS, reportez-vous au guide d'administration OneFS.

# Présentation de OneFS

OneFS regroupe les trois couches des architectures de stockage traditionnel (système de fichiers, gestionnaire de volume et protection des données) en une seule couche logicielle unifiée, constituant ainsi un seul système de fichiers intelligent et distribué qui s'exécute sur un cluster de stockage OneFS.


Figure 1 : OneFS associe un système de fichiers, un gestionnaire de volume et la protection des données dans un système unique, intelligent et distribué.

Il s'agit de l'innovation principale qui permet directement aux entreprises d'utiliser avec succès le stockage scale-out NAS dans leurs environnements actuels. OneFS est conforme aux principes clés du scale-out ; logiciel intelligent, composants matériels génériques et architecture distribuée. OneFS n'est pas uniquement le système d'exploitation, mais également le système de fichiers sous-jacent qui pilote et stocke les données dans le cluster.

## Nœuds PowerScale

OneFS fonctionne exclusivement avec les nœuds de plate-forme dédiés, appelés « cluster ». Un cluster unique se compose de plusieurs nœuds, qui sont des appliances d'entreprise montables en rack contenant la mémoire, le processeur, les composants de mise en réseau, l'interconnexion Ethernet ou InfiniBand faible latence, les contrôleurs de disque et les supports de stockage. Par conséquent, chaque nœud du cluster distribué possède des fonctions de calcul ainsi que des fonctions de stockage ou de capacité.

Avec l'architecture Gen 6, un seul boîtier de 4 nœuds dans un format 4RU (unités de rack) est nécessaire pour créer un cluster, qui peut évoluer jusqu'à 252 nœuds dans OneFS 8.2 et les versions ultérieures. Les plates-formes de nœuds individuelles nécessitent un minimum de trois nœuds et un espace de rack de 3RU pour former un cluster. Il existe plusieurs types de nœuds qui peuvent tous être intégrés dans un seul cluster sur lequel différents nœuds fournissent divers rapports de capacité de débit ou d'E/S par seconde (IOPS). Le boîtier de Gen 6 traditionnel et les nœuds autonomes PowerScale All-Flash F900, F600 et F200 coexisteront dans le même cluster.

Chaque nœud ou châssis ajouté à un cluster augmente la capacité cumulée en termes d'espace disque, de cache, de processeur et de réseau. OneFS valorise chaque blocs de construction matériel, ce qui signifie que l'ensemble est plus important que la somme des parties. La RAM est regroupée dans un seul cache cohérent, permettant ainsi aux E/S sur n'importe quelle partie du cluster de bénéficier des données mises en cache en tout lieu. Le journal du système de fichiers garantit que les écritures sont en sécurité lors des pannes d'alimentation. Les piles de disques et le CPU sont combinés pour augmenter le débit, la capacité et les E/S par seconde au fil de la croissance du cluster, pour l'accès à un seul fichier comme à plusieurs fichiers. La capacité de stockage d'un cluster peut varier de quelques dizaines de téraoctets à des dizaines de pétaoctets. La capacité maximale continuera d'évoluer à mesure que les supports de stockage et le châssis du nœud gagneront en densité.

Les plates-formes OneFS se divisent en plusieurs catégories, ou niveaux, suivant leur fonction :

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

Table 1 : Types de nœuds et niveaux matériels

## Réseau

Il existe deux types de réseaux associés à un cluster : interne et externe.

**Réseau back-end**

Toutes les communications intranœuds d'un cluster sont effectuées sur un réseau back-end dédié, comprenant une technologie 10, 40 ou 100 Gbit Ethernet, ou QDR InfiniBand (IB) à faible latence. Ce réseau back-end, qui est configuré avec des commutateurs redondants pour la haute disponibilité, agit en tant que backplane du cluster. Cela permet à chaque nœud d'agir en tant que contributeur du cluster et d'isoler la communication entre les nœuds sur un réseau privé, à haut débit et à faible latence. Ce réseau back-end utilise le protocole Internet (IP) pour la communication entre les nœuds.

**Réseau front-end**

Les clients se connectent au cluster à l'aide de connexions Ethernet (10 GbE, 25 GbE, 40 GbE ou 100 GbE) disponibles sur tous les nœuds. Puisque chaque nœud fournit ses propres ports Ethernet, la bande passante réseau disponible pour le cluster évolue de façon linéaire par rapport aux performances et à la capacité. Le cluster prend en charge les protocoles de communication réseau standard vers un réseau du client, notamment NFS, SMB, HTTP, FTP, HDFS et S3. En outre, OneFS fournit une intégration complète avec les environnements IPv4 et IPv6.

**Vue du cluster complet**

Le cluster complet est combiné avec le matériel, les logiciels, les réseaux dans la vue suivante :

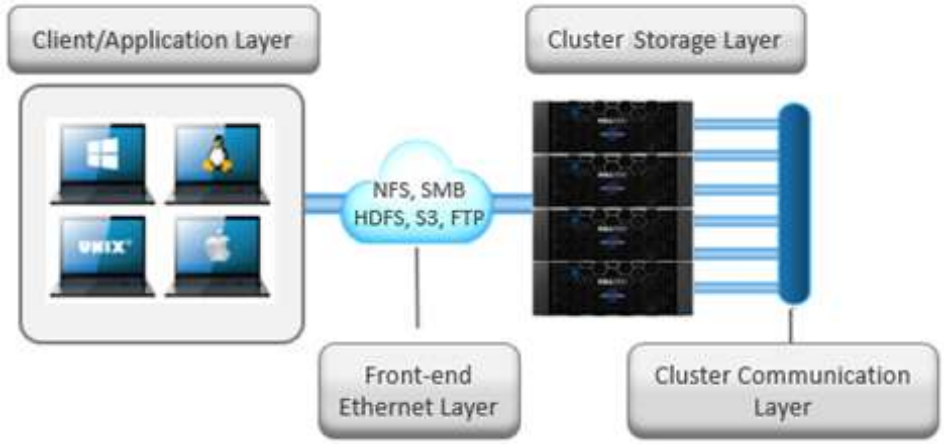


Figure 2 : Fonctionnement de tous les composants OneFS

Le diagramme ci-dessus illustre l'architecture complète : logiciels, matériel et réseau fonctionnant ensemble dans votre environnement ; les serveurs fournissant un système de fichiers unique totalement distribué qui peut évoluer de façon dynamique à mesure que les charges applicatives et les besoins en capacité ou en débit évoluent dans un environnement scale-out.

OneFS SmartConnect est un répartiteur de charge qui fonctionne sur la couche Ethernet front-end pour distribuer équitablement les connexions client sur l'ensemble du cluster. SmartConnect prend en charge le basculement sur incident NFS dynamique et le retour arrière pour les clients Linux et UNIX, ainsi que la disponibilité continue SMB3 pour les clients Windows. Cette technique garantit, en cas de défaillance de nœud ou d'opération de maintenance préventive, le transfert de toutes les opérations de lecture/écriture en cours vers un autre nœud du cluster afin de terminer l'opération sans perturber les utilisateurs ni les applications.

Lors d'un basculement sur incident, les clients sont répartis uniformément sur tous les nœuds restants du cluster, pour un impact minimal sur les performances. Si un nœud est mis hors ligne pour une raison quelconque, y compris pour une défaillance, les adresses IP virtuelles sur ce nœud sont migrées en toute transparence vers un autre nœud du cluster. Lorsque le nœud hors ligne est remis en ligne, SmartConnect rééquilibre automatiquement les clients NFS et SMB3 sur l'ensemble du cluster afin de garantir une utilisation optimale du stockage et des performances. Dans le cadre des mises à jour logicielles et de la maintenance périodique du système, cette fonctionnalité permet de procéder à des mises à niveau consécutives par nœud, pour une disponibilité complète pendant toute la durée de la fenêtre de maintenance.

📖 Pour plus d'informations, reportez-vous au livre blanc OneFS SmartConnect.

# Présentation du logiciel OneFS

### Système d'exploitation

OneFS est basé sur un système d'exploitation UNIX reposant sur BSD. Il prend en charge la sémantique Linux/UNIX et Windows en mode natif, y compris les liens matériels, l'effacement lors de la fermeture, la modification de noms atomique, les ACL et les attributs étendus. OneFS utilise BSD comme système d'exploitation de base, car il s'agit d'un système d'exploitation avancé et éprouvé. De plus, la communauté Open Source peut être utilisée pour l'innovation. À partir de OneFS 8.2, la version du système d'exploitation sous-jacente est FreeBSD 11.

### Services clients

Les protocoles front-end que les clients peuvent utiliser pour interagir avec OneFS sont désignés sous le nom de services clients. Reportez-vous à la section Protocoles pris en charge pour obtenir la liste détaillée des protocoles pris en charge. Pour mieux comprendre comment OneFS communique avec les clients, nous avons divisé le sous-système d'E/S en deux : la moitié supérieure, ou « initiateur », et la moitié inférieure, ou « participant ». Tous les nœuds du cluster sont des participants à une opération d'E/S spécifique. Le nœud auquel le client se connecte est l'initiateur, et ce nœud agit en tant que « capitaine » pour l'ensemble des opérations d'E/S. Les opérations de lecture et d'écriture sont détaillées dans les sections ultérieures.

### Opérations du cluster

Dans une architecture en clusters, il existe des tâches responsables de la santé et de la maintenance du cluster, ces tâches étant toutes gérées par le moteur de tâches OneFS. Le moteur de tâches s'exécute sur l'ensemble du cluster, et il est chargé de diviser et de maîtriser les tâches étendues de gestion et de protection du stockage. Pour y parvenir, il réduit une tâche en éléments de travail de petite taille, puis alloue ou mappe ces parties de la tâche globale à plusieurs threads de travail sur chaque nœud. La progression fait l'objet d'un suivi tout au long de l'exécution de la tâche, puis un rapport détaillé et un état sont présentés lorsqu'elle est terminée.

Le moteur de tâches inclut un système complet de point de contrôle qui permet de suspendre et de reprendre les tâches, en plus de les arrêter et de les démarrer. Le framework du moteur de tâches comprend également un système de gestion d'impact évolutif.

Le moteur de tâches exécute généralement les tâches en arrière-plan sur le cluster, à l'aide de capacités et de ressources de secours ou réservées à cet effet. Les tâches elles-mêmes peuvent être classées dans trois catégories principales :

### Tâches de maintenance du système de fichiers

Ces tâches assurent la maintenance du système de fichiers en arrière-plan et nécessitent généralement l'accès à tous les nœuds. Elles sont exécutées dans les configurations par défaut et, souvent, lorsque l'état du cluster est dégradé. Citons notamment la protection des systèmes de fichiers et les reconstructions de disque.

### Tâches de prise en charge de fonctions

Les tâches de prise en charge des fonctions effectuent un travail qui facilite certaines fonctions de gestion du stockage étendues et ne fonctionnent généralement que lorsque la fonction a été configurée. Il s'agit notamment de la déduplication et de l'analyse antivirus.

### Tâches d'action utilisateur

Ces tâches sont réalisées directement par l'administrateur de stockage afin d'atteindre un objectif de gestion des données particulier. Citons notamment la maintenance parallèle des suppressions et des autorisations liées aux arborescences.

Le tableau ci-dessous fournit la liste complète des tâches exposées du moteur de tâches, des opérations qu'elles effectuent et de leurs méthodes d'accès à leur système de fichiers respectif :

| Nom de la tâche | Description de la tâche | Méthode d'accès |
|---|---|---|
| **AutoBalance** | Équilibre l'espace disponible dans le cluster. | **Disque + LIN** |
| **AutoBalanceLin** | Équilibre l'espace disponible dans le cluster. | **LIN** |

| Nom de la tâche | Description de la tâche | Méthode d'accès |
|---|---|---|
| **AVScan** | Tâche d'analyse antivirus exécutée par le ou les serveurs antivirus. | **Arborescence** |
| **ChangelistCreate** | Crée une liste de modifications entre deux snapshots SyncIQ consécutifs | **Liste de modifications** |
| **CloudPoolsLin** | Archive les données vers un fournisseur de Cloud en fonction d'une règle de pools de fichiers. | **LIN** |
| **CloudPoolsTreewalk** | Archive les données vers un fournisseur de Cloud en fonction d'une règle de pools de fichiers. | **Arborescence** |
| **Collect** | Récupère l'espace disque qui n'a pas pu être libéré en raison d'un nœud ou d'un disque non disponible alors qu'il présente différentes conditions de panne. | **Disque + LIN** |
| **ComplianceStoreDelete** | Tâche de récupération d'espace en mode de conformité SmartLock. | **Arborescence** |
| **Dedupe** | Déduplique des blocs identiques du système de fichiers. | **Arborescence** |
| **DedupeAssessment** | Évalue en mode test les avantages de la déduplication. | **Arborescence** |
| **DomainMark** | Associe un chemin et son contenu à un domaine. | **Arborescence** |
| **DomainTag** | Associe un chemin et son contenu à un domaine. | **Arborescence** |
| **EsrsMftDownload** | Tâche de transfert de fichiers gérée par ESRS pour les fichiers de licence. | |
| **FilePolicy** | Tâche efficace de règle de pool de fichiers SmartPools. | **Liste de modifications** |
| **FlexProtect** | Reconstruit et rétablit la protection du système de fichiers après une panne. | **Disque + LIN** |
| **FlexProtectLin** | Rétablit la protection du système de fichiers. | **LIN** |
| **FSAnalyze** | Collecte des données d'analyse de système de fichiers qui sont utilisées avec InsightIQ. | **Liste de modifications** |
| **IndexUpdate** | Crée et met à jour un index de système de fichiers efficace pour les tâches FilePolicy et FSAnalyze, | **Liste de modifications** |
| **IntegrityScan** | Procède à la vérification et à la correction en ligne des éventuelles incohérences de système de fichiers. | **LIN** |
| **LinCount** | Analyse et comptabilise les inodes logiques (LIN) du système de fichiers. | **LIN** |
| **MediaScan** | Analyse les disques afin de détecter les éventuelles erreurs au niveau des médias. | **Disque + LIN** |
| **MultiScan** | Exécute les tâches Collect et AutoBalance simultanément. | **LIN** |
| **PermissionRepair** | Autorisations adéquates sur les fichiers et répertoires. | **Arborescence** |
| **QuotaScan** | Met à jour la comptabilité des quotas pour les domaines créés sur un chemin de répertoire existant. | **Arborescence** |
| **SetProtectPlus** | Applique la règle de fichier par défaut. Cette tâche est désactivée si SmartPools est activé sur le cluster. | **LIN** |

| Nom de la tâche | Description de la tâche | Méthode d'accès |
|---|---|---|
| **ShadowStoreDelete** | Libère l'espace associé à une zone de stockage de clichés instantanés. | **LIN** |
| **ShadowStoreProtect** | Protège les zones de stockage de clichés instantanés qui sont référencées par un LIN avec une protection demandée plus élevée. | **LIN** |
| **ShadowStoreRepair** | Répare les zones de stockage de clichés instantanés. | **LIN** |
| **SmartPools** | Tâche qui déplace les données entre les niveaux des nœuds d'un cluster. Exécute également la fonctionnalité CloudPools si elle est activée par une licence et configurée. | **LIN** |
| **SmartPoolsTree** | Applique les règles de fichier SmartPools dans une sous-arborescence. | **Arborescence** |
| **SnapRevert** | Inverse l'intégralité d'un snapshot. | **LIN** |
| **SnapshotDelete** | Libère l'espace disque associé aux snapshots supprimés. | **LIN** |
| **TreeDelete** | Supprime un chemin d'accès du système de fichiers directement à partir du cluster lui-même. | **Arborescence** |
| **Undedupe** | Supprime la déduplication de blocs identiques dans le système de fichiers. | **Arborescence** |
| **Mettre à niveau de** | Met à niveau le cluster sur une version ultérieure de OneFS. | **Arborescence** |
| **WormQueue** | Analyse la file d'attente SmartLock LIN | **LIN** |

Figure 1 : Descriptions des tâches du moteur de tâches OneFS

Bien que les tâches de maintenance du système de fichiers soient exécutées par défaut, à la suite d'une planification ou en réponse à un événement spécifique du système de fichiers, toutes les tâches du moteur de tâches peuvent être gérées par le biais de la configuration de leur niveau de priorité (par rapport aux autres tâches) et de leur règle d'impact.

Une règle d'impact peut comporter un ou plusieurs intervalles d'impact, qui définissent des périodes sur une semaine donnée. Chaque intervalle d'impact peut être configuré pour utiliser un seul niveau d'impact prédéfini qui spécifie la quantité de ressources de cluster à utiliser pour une opération de cluster particulière. Les niveaux d'impact disponibles du moteur de tâches sont les suivants :

- Paused
- Basse
- Moyenne
- High

Ce degré de granularité permet de configurer les intervalles et les niveaux d'impact tâche par tâche, afin de garantir un fonctionnement fluide du cluster. Les règles d'impact qui en résultent déterminent le moment d'exécution d'une tâche et les ressources qu'elle peut consommer.

En outre, les tâches du moteur de tâches sont hiérarchisées sur une échelle de un à dix. Plus la valeur est faible, plus la priorité est élevée. Ce concept est semblable au concept de l'utilitaire de planification UNIX, « nice ».

Le moteur de tâches OneFS permet l'exécution simultanée d'un maximum de trois tâches. Cette exécution simultanée est régie par les critères suivants :

- Priorité des tâches
- Jeu d'exclusions : tâches qui ne peuvent pas s'exécuter ensemble (FlexProtect et AutoBalance, par exemple)
- Intégrité du cluster : la plupart des tâches ne peuvent pas s'exécuter lorsque le cluster a l'état dégradé.

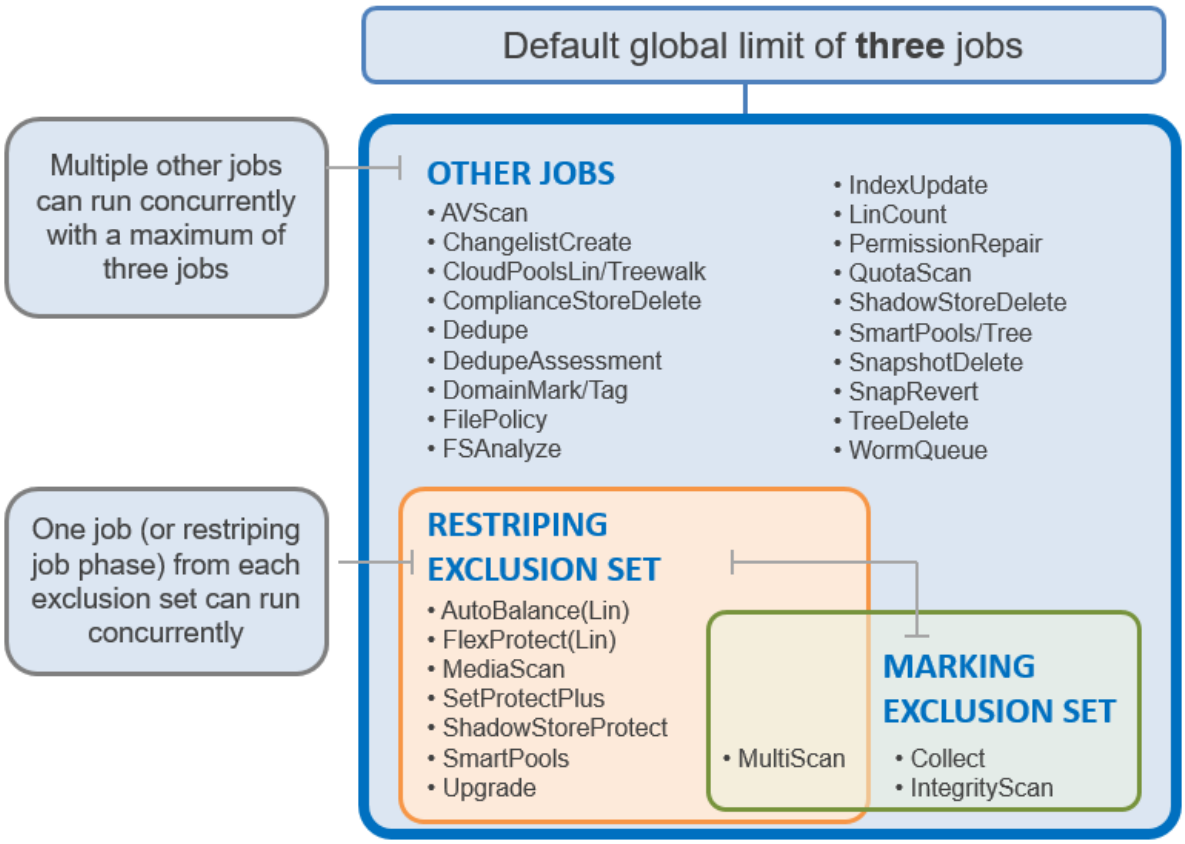


Figure 4 : jeux d'exclusion du moteur de tâches OneFS

🕮 Pour plus d'informations, reportez-vous au livre blanc OneFS Job Engine.

# Structure du système de fichiers

Le système de fichiers OneFS est basé sur le système de fichiers Unix (UFS) et, par conséquent, est un système de fichiers DFS très rapide. Chaque cluster crée un espace de nommage et un système de fichiers uniques. Cela signifie que le système de fichiers est distribué sur tous les nœuds du cluster et est accessible par les clients qui se connectent à n'importe quel nœud du cluster. Il n'y a aucun partitionnement et il n'est pas nécessaire de créer des volumes. Au lieu de limiter l'accès à l'espace disponible et à des fichiers non autorisés au niveau du volume physique, OneFS fournit les mêmes fonctions logicielles par le biais du partage et des autorisations de fichier, et par le biais du service SmartQuotas, qui vous permet de gérer les quotas au niveau du répertoire.

🕮 Pour plus d'informations, reportez-vous au livre blanc OneFS SmartQuotas.

Toutes les informations étant partagées entre les nœuds sur le réseau interne, les données peuvent être écrites et lues à partir de n'importe quel nœud, ce qui optimise les performances lorsque plusieurs utilisateurs lisent et écrivent simultanément sur le même jeu de données.

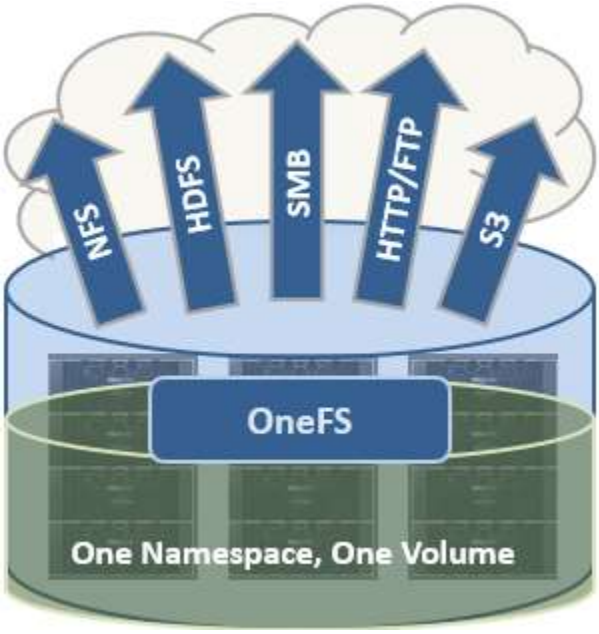


Figure 5 : Système de fichiers unique avec plusieurs protocoles d'accès

OneFS est un système de fichiers unique doté d'un espace de nommage. Les données et les métadonnées sont réparties entre les nœuds à des fins de redondance et de disponibilité. Le stockage a été entièrement virtualisé pour les utilisateurs et l'administrateur. L'arborescence de fichiers peut augmenter de façon organique sans qu'il soit nécessaire de planifier ni de superviser le mode de croissance de l'arborescence ou la manière dont les utilisateurs l'emploient. L'administrateur n'a pas à se préoccuper de la hiérarchisation des fichiers sur le disque approprié, car OneFS SmartPools s'en charge automatiquement sans générer d'interruption pour l'arborescence unique. Il n'est pas nécessaire d'accorder une attention particulière à la méthode de réplication d'une telle arborescence, car le service OneFS SyncIQ parallélise automatiquement le transfert de l'arborescence de fichiers vers un ou plusieurs autres clusters, sans tenir compte de la forme ou de la profondeur de l'arborescence de fichiers.

Cette conception doit être comparée à l'agrégation d'espace de nommage, une technologie couramment utilisée pour que le NAS traditionnel « semble » disposer d'un espace de nommage unique. Avec l'agrégation des espaces de nommage, les fichiers doivent quand même être gérés dans des volumes distincts, mais une simple couche de « placage » permet de « coller » différents répertoires de volumes à un niveau supérieur de l'arborescence via des liens symboliques. Dans ce modèle, les LUN et les volumes, ainsi que les limites de volume, sont toujours présents. Les fichiers doivent être déplacés manuellement d'un volume à l'autre à des fins d'équilibrage de la charge. L'administrateur doit bien réfléchir à l'organisation de l'arborescence. La hiérarchisation est loin d'être transparente et nécessite une intervention importante et continue. Le basculement sur incident requiert la mise en miroir des fichiers entre les volumes, ce qui réduit l'efficacité et accroît les coûts d'achat, d'alimentation et de ventilation. D'une manière générale, lors de l'utilisation de l'agrégation des espaces de nommage, la charge de l'administrateur est plus élevée que pour un périphérique NAS traditionnel simple. Cela évite à ces infrastructures de devenir très volumineuses.

## Répartition des données

OneFS utilise des pointeurs physiques et des extensions pour les métadonnées et stocke les métadonnées des fichiers et des répertoires dans les inodes. Les inodes logiques OneFS (LIN) font généralement 512 octets, ce qui leur permet de s'adapter aux secteurs natifs avec lesquels la majorité des disques durs sont formatés. Il est également possible de prendre en charge les inodes de 8 Ko, afin de prendre en charge les classes de disques durs plus denses qui sont désormais formatées avec des secteurs de 4 Ko.

Les arborescences B sont largement utilisées dans le système de fichiers, ce qui autorise l'évolutivité vers des milliards d'objets et des recherches quasi instantanées de données ou de métadonnées. OneFS est un système de fichiers hautement distribué et complètement symétrique. Les données et les métadonnées sont toujours redondantes sur plusieurs périphériques matériels. Les données sont protégées à l'aide du codage d'effacement sur les nœuds du cluster. Cela crée un cluster haute efficacité autorisant un ratio capacité brute-capacité utile de 80 % ou plus sur les clusters de cinq nœuds ou plus. Les métadonnées (qui composent généralement moins de 1 % du système) sont mises en miroir dans le cluster pour accroître les performances et la disponibilité. Étant donné que OneFS ne dépend pas de la technologie RAID, la capacité de redondance est définie par l'administrateur, au niveau fichier ou répertoire, au-dessus des valeurs par défaut du cluster. L'accès aux métadonnées et le verrouillage sont gérés par tous les nœuds de façon collective et égale dans une architecture peer-to-peer. Cette symétrie est essentielle à la simplicité et à la résilience de l'architecture. Il n'y a aucun serveur de métadonnées, gestionnaire de verrou ou nœud de passerelle.

Puisque OneFS doit accéder à des blocs à partir de plusieurs périphériques au même moment, le système d'adressage utilisé pour les données et les métadonnées est indexé au niveau physique par un tuple de {nœud, disque, décalage}. Par exemple, si 12345 est l'adresse d'un bloc situé sur le disque 2 du nœud 3, il faut lire {3,2,12345}. Toutes les métadonnées du cluster sont mises en miroir plusieurs fois à des fins de protection des données, au minimum au niveau de redondance du fichier associé. Par exemple, si un fichier dispose d'une protection par code d'effacement « +2n », son utilisation pourrait résister à deux défaillances simultanées. Ensuite, toutes les métadonnées nécessaires pour accéder à ce fichier seraient mises en miroir trois fois afin que ce dernier puisse résister à deux défaillances. Le système de fichiers permet de façon inhérente à n'importe quelle structure d'utiliser n'importe quel bloc (ou l'intégralité des blocs), sur n'importe quel nœud du cluster.

Les autres systèmes de stockage envoient les données via les couches RAID et de gestion des volumes, d'où un manque d'efficacité au niveau de la répartition des données et un accès non optimisé en mode bloc. OneFS gère directement le positionnement des fichiers, jusqu'au niveau du secteur, sur n'importe quel disque du cluster. Cela permet d'optimiser le positionnement des données et les schémas d'E/S, et évite les opérations inutiles de lecture-modification-écriture. En plaçant les données sur les disques fichier par fichier, OneFS peut gérer avec flexibilité le type de répartition ainsi que la redondance du système de stockage au niveau du système, du répertoire ou même des fichiers. Les systèmes de stockage traditionnels nécessiteraient que tout un volume RAID soit dédié à un type de performances et à un paramètre de protection particuliers. Par exemple, un ensemble de disques pourrait être réorganisé dans une protection RAID 1+0 pour une base de données. Il est donc difficile d'optimiser l'utilisation de la pile de disques sur l'ensemble du domaine de stockage (car les piles inactives ne peuvent pas être empruntées). Cela mène à des conceptions rigides qui ne sont pas adaptées aux besoins métiers. OneFS autorise des réglages individuels et des modifications flexibles à tout moment et en ligne.

## Écritures de fichier

Le logiciel OneFS s'exécute de façon égale sur tous les nœuds, ce qui crée un système de fichiers unique qui s'exécute sur chaque nœud. Aucun nœud ne contrôle le cluster ; tous les nœuds sont de véritables homologues.

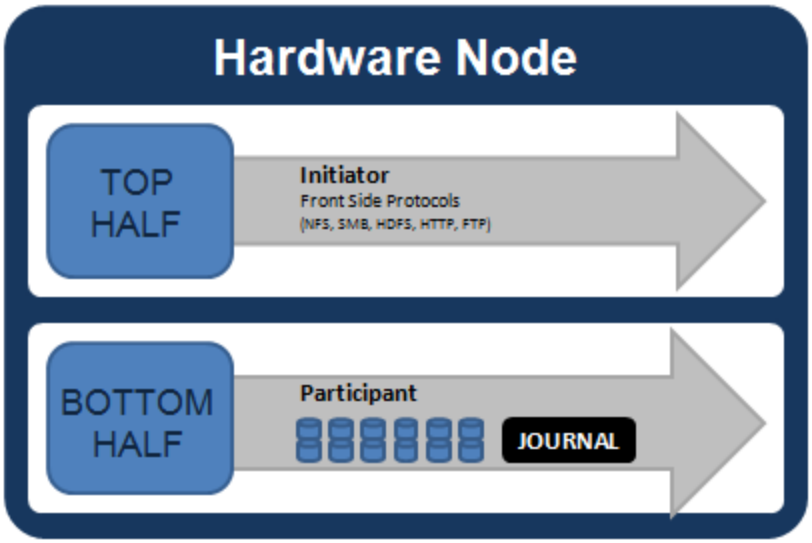


Figure 6 : Modèle de composants de nœud impliqués dans les E/S

Si nous devions visualiser tous les composants de chaque nœud d'un cluster impliqués dans les E/S d'un point de vue général, cela ressemblerait à la figure 6 ci-dessus. Nous avons divisé la pile en une couche supérieure, appelée initiateur, et une couche inférieure, appelée participant. Cette division est utilisée comme « modèle logique » pour l'analyse des données en lecture ou écriture. À un niveau physique, les CPU et le cache RAM des nœuds gèrent simultanément les tâches de l'initiateur et du participant pour les E/S qui se produisent dans l'ensemble du cluster. Certains caches et un gestionnaire de verrou distribué sont exclus du schéma ci-dessus pour en conserver la simplicité. Ils sont abordés dans les sections ultérieures de ce document.

Lorsqu'un client se connecte à un nœud pour écrire dans un fichier, il se connecte à la partie supérieure, ou initiateur, de ce nœud. Les fichiers sont décomposés en petits fragments logiques nommés bandes avant d'être écrits dans la moitié inférieure, ou participant, d'un nœud (disque). La mise en mémoire tampon en toute sécurité à l'aide d'un fusionneur d'écriture permet de s'assurer que les écritures sont efficaces et que les opérations de lecture-modification-écriture sont évitées. La taille de chaque fragment de fichier est appelée taille d'unité de bande.

OneFS répartit les données sur tous les nœuds, et pas seulement entre les disques, et protège les fichiers, les répertoires et les métadonnées associées via le codage d'effacement logiciel ou la technologie de mise en miroir. Pour les données, OneFS peut utiliser (à la seule discrétion de l'administrateur), le système de codage d'effacement Reed-Solomon pour assurer la protection des données, ou la mise en miroir (moins fréquente). La mise en miroir, lorsqu'elle est appliquée aux données utilisateur, a tendance à être plus utilisée lorsque les performances impliquent des taux élevés de transactions. La majeure partie des données utilisateur utilise habituellement le codage d'effacement, car il fournit de très hautes performances sans sacrifier l'efficacité sur disque. Le codage d'effacement peut fournir une efficacité supérieure à 80 % sur un disque brut de cinq nœuds ou plus et, sur de grands clusters, peut faire de même tout en fournissant une redondance quatre fois plus importante. La largeur de répartition de n'importe quel fichier correspond au nombre de nœuds (et non de disques) sur lequel un fichier est écrit. Elle est déterminée par le nombre de nœuds du cluster, la taille du fichier et le paramètre de protection (par exemple, +2n).

OneFS utilise des algorithmes avancés pour déterminer la répartition des données qui offrira les performances et l'efficacité maximales. Lorsqu'un client se connecte à un nœud, l'initiateur de ce nœud agit en tant que « capitaine » pour la répartition des données d'écriture de ce fichier. La protection par code d'effacement (ECC), les données, les métadonnées et les inodes sont tous distribués sur plusieurs nœuds dans un cluster, voire plusieurs disques dans les nœuds.

La figure 7 ci-dessous illustre une écriture de fichier sur l'ensemble des nœuds d'un cluster à trois nœuds.

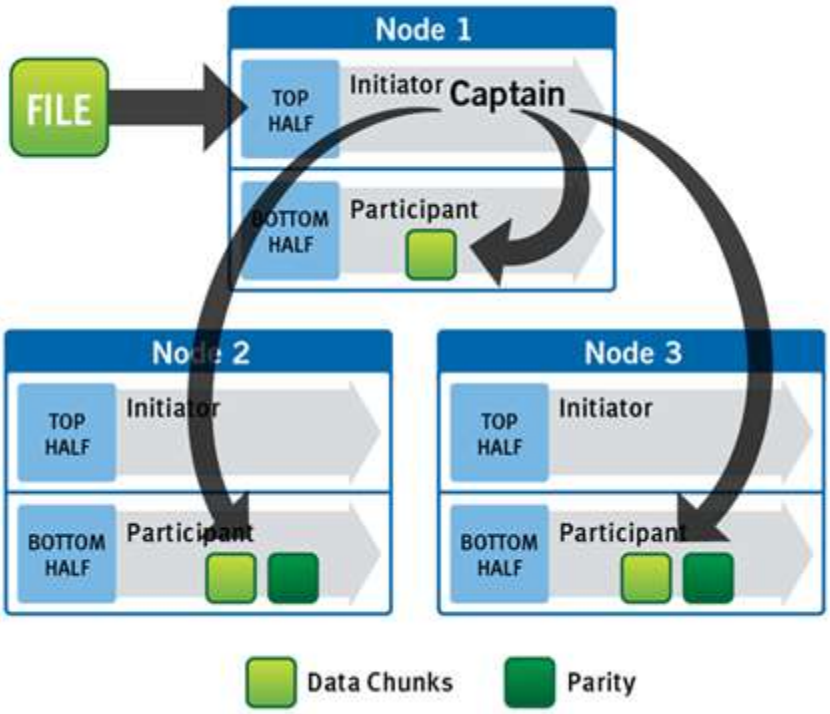


Figure 7 : Opération d'écriture de fichier sur un cluster à trois nœuds

OneFS utilise le réseau back-end pour allouer et répartir les données sur tous les nœuds du cluster automatiquement ; par conséquent, aucun traitement supplémentaire n'est requis. À mesure que les données sont écrites, elles sont protégées au niveau spécifié. Lorsque les opérations d'écriture sont effectuées, OneFS divise les données en unités atomiques appelées groupes de protection. La redondance est intégrée dans des groupes de protection, de telle sorte que si chaque groupe de protection est sécurisé, le fichier entier est sécurisé. Pour les fichiers protégés par des codes d'effacement, un groupe de protection se compose d'une série de blocs de données, ainsi que d'un ensemble de codes d'effacement pour ces blocs de données. Pour les fichiers mis en miroir, un groupe de protection comprend tous les miroirs d'un ensemble de blocs. OneFS est capable de changer le type du groupe de protection utilisé dans un fichier de façon dynamique, à mesure de l'écriture. Cela peut autoriser de nombreuses fonctions supplémentaires, notamment en permettant au système de continuer sans blocage dans les situations où les défaillances temporaires de nœud dans le cluster empêcheraient d'utiliser le nombre souhaité de codes d'effacement. La mise en miroir peut être utilisée temporairement dans ces cas pour permettre aux écritures de continuer. Lorsque les nœuds sont restaurés dans le cluster, ces groupes de protection mis en miroir sont convertis de manière transparente et automatique en code d'effacement protégé, sans intervention de l'administrateur.

La taille de bloc du système de fichiers OneFS est de 8 Ko. Un fichier dont la taille est inférieure à 8 Ko utilisera un bloc de 8 Ko complet. En fonction du niveau de protection des données, ce fichier de 8 Ko pourrait utiliser plus de 8 Ko d'espace de données. Notez que les paramètres de protection des données sont décrits en détail dans une section ultérieure du présent document. OneFS peut prendre en charge les systèmes de fichiers comptant des milliards de petits fichiers avec de très hautes performances, car toutes les structures sur disque sont conçues pour s'adapter à de telles tailles et fournissent un accès quasi instantané à n'importe quel objet, quel que soit le nombre total d'objets. Pour les fichiers volumineux, OneFS peut utiliser plusieurs blocs de 8 Ko consécutifs. Dans ce cas, jusqu'à seize blocs contigus peuvent être agrégés par bandes sur un disque à un seul nœud. Dans le cas d'un fichier de 32 Ko, quatre blocs contigus de 8 Ko seront utilisés.

Pour les fichiers encore plus volumineux, OneFS peut optimiser les performances séquentielles en tirant parti d'une unité de bande composée de 16 blocs contigus, pour un total de 128 Ko par unité de bande. Lors d'une opération d'écriture, les données sont divisées en unités de bande elles-mêmes réparties sur plusieurs nœuds en tant que groupes de protection. À mesure que les données sont placées dans le cluster, les codes d'effacement ou les miroirs, en fonction des besoins, sont distribués au sein de chaque groupe de protection pour assurer la protection des fichiers à tout moment.

La fonction AutoBalance de OneFS permet, entre autres, de réallouer et de rééquilibrer automatiquement les données et de rendre l'espace de stockage plus utile et plus efficace lorsque cela est possible. Dans la plupart des cas, la largeur de bande des fichiers volumineux peut être augmentée pour tirer parti du nouvel espace disponible (à mesure que vous ajoutez des nœuds), afin que la répartition sur disque soit plus efficace. AutoBalance maintient une efficacité élevée sur disque et élimine automatiquement les points sensibles.

La partie supérieure de l'initiateur du nœud « capitaine » utilise une transaction de validation à deux phases, modifiée pour distribuer en toute sécurité les écritures à plusieurs NVRAM sur l'ensemble du cluster, comme illustré à la figure 8 ci-dessous.

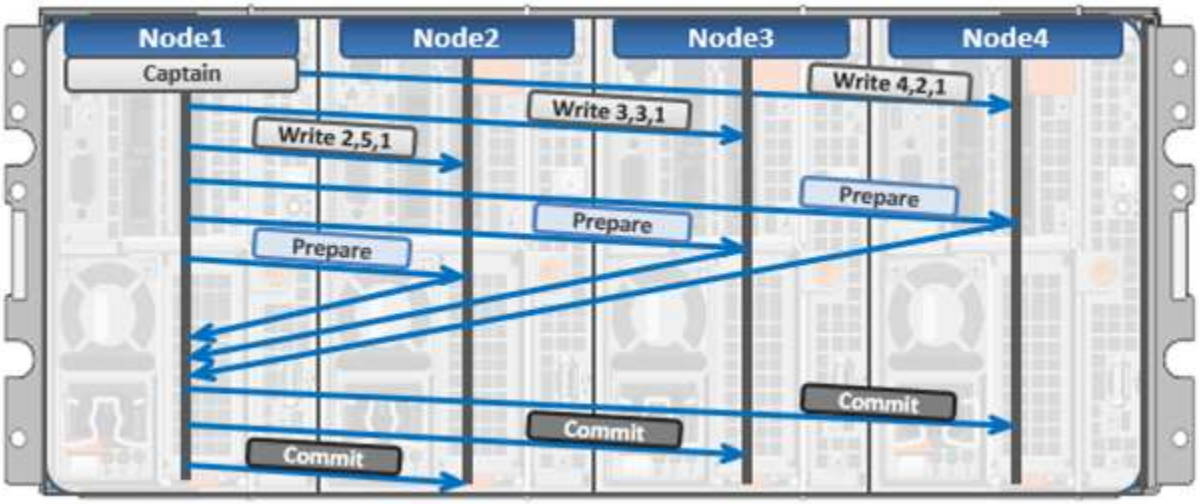


Figure 8 : Transactions distribuées avec validation à deux phases

Chaque nœud détenant des blocs dans une écriture particulière est impliqué dans une validation à deux phases. Le mécanisme s'appuie sur la NVRAM pour consigner toutes les transactions effectuées sur chaque nœud du cluster de stockage. L'utilisation de plusieurs NVRAM en parallèle autorise les écritures à haut débit tout en préservant la sécurité de données contre tous les types de défaillance, y compris les coupures d'alimentation. Si un nœud tombe en panne au cours d'une transaction, la transaction est redémarrée instantanément sans ce nœud. Lorsque le nœud est rétabli, les seules actions requises sont la relecture de son journal depuis la NVRAM, qui prend quelques secondes ou minutes, et le rééquilibrage par AutoBalance des fichiers impliqués dans la transaction. Les processus coûteux « fsck » ou « disk-check » ne sont jamais obligatoires. Aucune longue resynchronisation n'est jamais nécessaire. Les écritures ne sont jamais bloquées à la suite d'une panne. Ce système de transaction breveté est l'une des méthodes utilisées par OneFS pour éliminer un seul point de défaillance, voire plusieurs.

Dans une opération d'écriture, l'initiateur (capitaine) dirige ou orchestre la répartition des données et des métadonnées, la création des codes d'effacement et le fonctionnement normal de gestion des verrous et de contrôle des autorisations. À partir de l'interface de gestion Web ou de l'interface CLI, un administrateur peut à tout moment optimiser les décisions de répartition prises par OneFS pour améliorer le workflow. L'administrateur peut effectuer un choix parmi les modèles d'accès ci-dessous, à un niveau de fichier individuel ou de répertoire :

- Concurrence : Optimise la charge actuelle sur le cluster, avec de nombreux clients simultanés. Ce paramètre constitue le meilleur comportement pour les charges applicatives mixtes.
- Streaming : Optimise la lecture en temps réel haut débit d'un seul fichier, par exemple pour activer la lecture très rapide avec un client unique.
- Aléatoire : Optimise l'accès imprévisible au fichier en ajustant l'agrégation par bandes et en désactivant l'utilisation de tous les caches de lecture préalable.

OneFS inclut également un mécanisme de préchargement évolutif en temps réel, afin de fournir des performances de lecture optimales pour les fichiers avec un modèle d'accès reconnu, sans aucune intervention de l'administrateur.

ⓘ La plus grande taille de fichier actuellement prise en charge par OneFS est augmentée à 16 To dans OneFS 8.2.2 et les versions ultérieures, contre un maximum de 4 To dans les versions précédentes.

# Mise en cache OneFS

La conception de l'infrastructure de mise en cache OneFS repose sur l'agrégation du cache présent sur chaque nœud d'un cluster dans un pool de mémoire accessible dans le monde entier. Pour ce faire, OneFS utilise un système de messagerie efficace, similaire à l'accès mémoire non uniforme (NUMA). Cela permet au cache de mémoire de tous les nœuds d'être disponible pour chacun des nœuds du cluster. La mémoire distante est accessible par le biais d'une interconnexion interne et sa latence est inférieure à celle des disques durs.

Pour l'accès distant à la mémoire, OneFS utilise un réseau Ethernet à plat redondant et en sous-réseaux, en tant que bus système distribué. Bien qu'il ne soit pas aussi rapide que l'accès à la mémoire locale, l'accès à la mémoire distante est tout de même très rapide grâce à la faible latence d'Ethernet de 40 GbE.

Le sous-système de mise en cache OneFS est cohérent sur l'ensemble du cluster. Cela signifie que si le même contenu existe dans les caches privés de plusieurs nœuds, les données mises en cache seront cohérentes sur toutes les instances. OneFS utilise le protocole MESI pour conserver la cohérence de cache. Ce protocole met en œuvre une règle d'invalidation lors de l'écriture pour garantir la cohérence de toutes les données sur l'ensemble du cache partagé.

OneFS utilise jusqu'à trois niveaux de cache de lecture, ainsi qu'un cache d'écriture s'appuyant sur la NVRAM, ou fusionneur. Ceux-ci, ainsi que leur interaction générale, sont illustrés sur le schéma suivant.

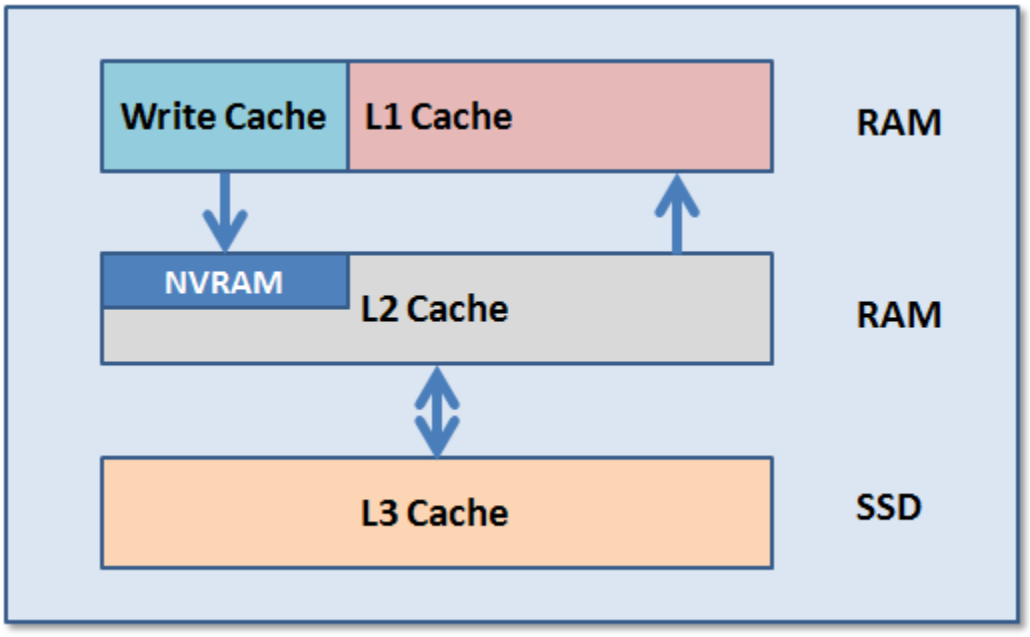


Figure 9 : Hiérarchie de mise en cache OneFS

Les deux premiers types de cache de lecture, de niveau 1 (N1) et de niveau 2 (N2), sont basés sur la mémoire (RAM) et similaires au cache utilisé dans les processeurs (CPU). Ces deux couches de cache sont présentes dans tous les nœuds de stockage de la plate-forme.

| Nom | Type | Persistance | Description |
|---|---|---|---|
| **Cache L1** | RAM | Volatile | Également appelé cache front-end, il contient les copies propres et cohérentes au niveau des clusters des blocs de données et de métadonnées des systèmes de fichiers demandés par l'intermédiaire de clients, par le biais du réseau front-end |
| **Cache L2** | RAM | Volatile | Cache back-end contenant les copies propres des données et métadonnées des systèmes de fichiers sur un nœud local. |
| **SmartCache / Fusionneur d'écritures** | NVRAM | Non volatile | Cache de journal NVRAM persistant et alimenté par une batterie qui met en mémoire tampon les écritures en attente vers des fichiers front-end qui n'ont pas été copiés sur disque. |
| **SmartFlash Cache L3** | Disque SSD | Non volatile | Contient des blocs de données et de métadonnées de fichiers exclus du cache N2, augmentant ainsi la capacité du cache N2. |

## Cohérence du cache OneFS

Le sous-système de mise en cache OneFS est cohérent sur l'ensemble du cluster. Cela signifie que si le même contenu existe dans les caches privés de plusieurs nœuds, les données mises en cache seront cohérentes sur toutes les instances. Prenez par exemple l'état initial et la séquence d'événements suivants :

1. Le nœud 1 et le nœud 5 disposent chacun d'une copie de données située à une adresse dans le cache partagé.
2. En réponse à une demande d'écriture, le nœud 5 invalide la copie du nœud 1.
3. Le nœud 5 met ensuite la valeur à jour. (Reportez-vous à la section ci-dessous.)
4. Le nœud 1 doit relire les données à partir du cache partagé pour obtenir la valeur mise à jour.

OneFS utilise le protocole MESI pour conserver la cohérence de cache. Ce protocole met en œuvre une règle d'invalidation lors de l'écriture pour garantir la cohérence de toutes les données sur l'ensemble du cache partagé. Le schéma suivant illustre les différents états possibles des données mises en cache, ainsi que les transitions entre ces états. Les différents états illustrés sont les suivants :

- M - Modifiées : Les données existent uniquement dans le cache local et ont été modifiées par rapport à la valeur dans le cache partagé. les données modifiées sont généralement définies comme étant non synchronisées.
- E - Exclusives : les données existent uniquement dans le cache local, mais correspondent à la valeur dans le cache partagé ; ces données sont souvent décrites comme propres.
- P - Partagées : les données d'un cache local peuvent également se trouver dans d'autres caches locaux du cluster.
- N - Non valide : un verrou (exclusif ou partagé) a été perdu sur les données.

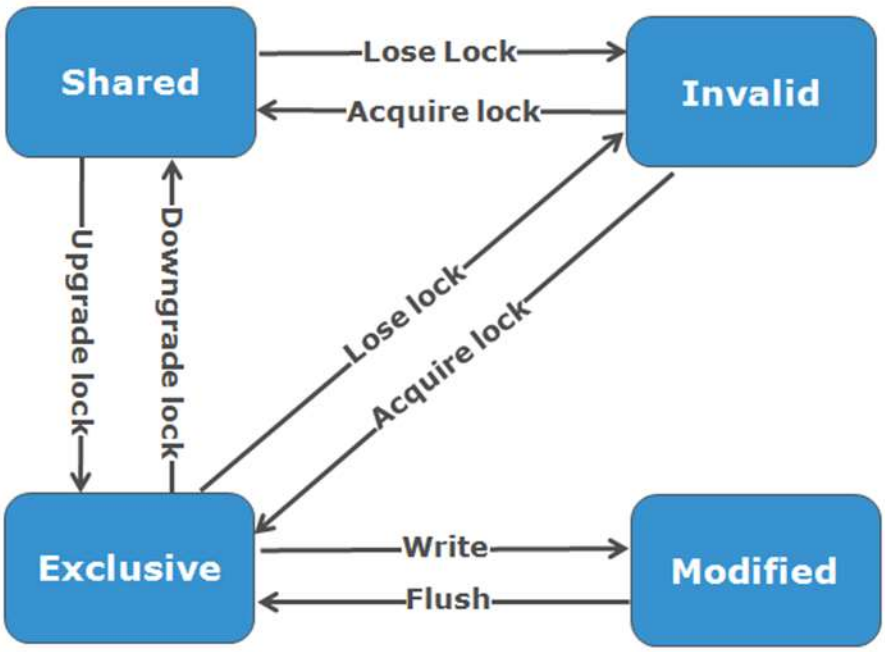


Figure 10 : Schéma des états de la cohérence du cache OneFS

## Cache de niveau 1

Le cache de niveau 1 (L1), ou cache front-end, est la mémoire la plus proche des couches de protocole (NFS, SMB, etc.) utilisées par les clients, ou initiateurs, connectés à ce nœud. Le principal objectif du cache N1 est d'effectuer une lecture préalable des données de nœuds distants. Les données font l'objet d'une lecture préalable par fichier et sont optimisées afin de réduire la latence associée au réseau back-end des nœuds. Étant donné que la latence de l'interconnexion backend est relativement faible, la taille du cache L1 et la quantité de données généralement stockées par demande sont inférieures à celles du cache L2.

Le cache N1 est également appelé cache distant car il contient des données récupérées dans d'autres nœuds du cluster. Il est cohérent sur l'ensemble du cluster, mais est uniquement utilisé par le nœud sur lequel il se trouve, les autres nœuds ne pouvant y accéder. Les données figurant dans le cache N1 des nœuds de stockage sont rejetées après utilisation. Le cache N1 utilise l'adressage basé sur des fichiers, qui permet d'accéder aux données via un décalage dans un objet de fichier.

Le cache N1 fait référence à la mémoire située sur le même nœud que l'initiateur. Uniquement accessible sur le nœud local, le cache n'est généralement pas la copie principale des données. Il est similaire au cache N1 sur un cœur de CPU, qui peut être invalidé lorsque d'autres cœurs écrivent sur la mémoire principale.

La cohérence du cache N1 est gérée via un protocole de type MESI à l'aide de verrous distribués, comme décrit ci-dessus.

OneFS utilise également un cache d'inodes dédié dans lequel les inodes récemment demandés sont conservés. Le cache d'inodes a souvent un impact considérable sur les performances, car les clients mettent fréquemment des données en cache, et de nombreuses activités d'E/S réseau sont principalement des demandes d'attributs et de métadonnées de fichiers, qui peuvent être rapidement renvoyés à partir de l'inode mis en cache.

ⓘ Le cache L1 est utilisé différemment dans les nœuds de cluster Accelerator, qui ne contiennent aucun lecteur de disque. Dans ce cas, l'ensemble du cache de lecture correspond au cache N1, car toutes les données sont extraites d'autres nœuds de stockage. En outre, l'obsolescence du cache se base sur une règle d'éviction LRU (Least Recently Used), et non sur l'algorithme d'abandon généralement utilisé dans le cache N1 d'un nœud de stockage. Étant donné que le cache N1 d'un accélérateur est volumineux et que les données qu'il contient sont particulièrement susceptibles de faire l'objet d'une nouvelle demande, les blocs de données ne sont pas immédiatement supprimés du cache après leur utilisation. Toutefois, les charges applicatives comprenant énormément de métadonnées ou faisant l'objet de nombreuses mises à jour n'en bénéficient pas autant. Le cache d'un accélérateur est uniquement bénéfique aux clients directement connectés au nœud.

## Cache de niveau 2

Le cache de niveau 2 (N2), ou cache back-end, fait référence à la mémoire locale située sur le nœud sur lequel un bloc de données spécifique est stocké. Le cache N2 est globalement accessible à partir de n'importe quel nœud du cluster et permet de réduire la latence d'une opération de lecture en évitant une recherche directement à partir des disques. Par conséquent, la quantité de données faisant l'objet d'une lecture préalable dans le cache N2 pour une utilisation par les nœuds distants est bien supérieure à celle dans le cache N1.

Le cache N2 est également appelé cache local car il contient des données récupérées à partir des disques situés sur ce nœud, puis mises à disposition pour les demandes à partir de nœuds distants. Les données du cache N2 sont exclues conformément à un algorithme LRU (Least Recently Used).

Les données du cache N2 sont traitées par le nœud local à l'aide d'un décalage dans un disque local de ce nœud. Étant donné que le nœud sait où les données demandées par les nœuds distants sont situées sur le disque, il s'agit d'un moyen très rapide de récupérer des données destinées à des nœuds distants. Un nœud distant accède au cache N2 en effectuant une recherche de l'adresse du bloc d'un objet de fichier particulier. Comme décrit ci-dessus, l'invalidation MESI n'est pas nécessaire ; le cache est mis à jour automatiquement lors des opérations d'écriture et sa cohérence est assurée par le système transactionnel et la NVRAM.

## Cache de niveau 3

Un troisième niveau facultatif de cache de lecture, appelé SmartFlash ou cache de niveau 3 (N3), peut également être configuré sur les nœuds qui contiennent des disques SSD. SmartFlash (N3) est un cache d'exclusion rempli par des blocs du cache N2 à mesure qu'ils deviennent obsolètes et sont retirés de la mémoire. Il existe plusieurs avantages liés à l'utilisation de disques SSD plutôt que de périphériques de stockage de système de fichiers traditionnels pour la mise en cache. Par exemple, lorsqu'il est réservé à la mise en cache, le disque SSD est entièrement utilisé, et les écritures se produisent de façon très linéaire et prévisible. Cela garantit un taux d'utilisation bien supérieur et se traduit également par une usure considérablement réduite et une durabilité largement accrue par rapport à l'utilisation d'un système de fichiers standard, en particulier pour les charges applicatives en écriture aléatoire. L'utilisation de disques SSD pour le cache rend également le dimensionnement de la capacité SSD beaucoup plus simple et moins source d'erreur par rapport à l'utilisation des disques SSD en tant que niveau de stockage.

Le schéma suivant illustre la façon dont les clients interagissent avec l'infrastructure de cache de lecture et le fusionneur d'écritures OneFS. Le cache N1 interagit toujours avec le cache N2 sur n'importe quel nœud qui le nécessite, et le cache N2 interagit à la fois avec le sous-système de stockage et le cache N3. Le cache N3 est stocké sur un disque SSD dans le nœud, et il est activé sur chaque nœud du même pool de nœuds.

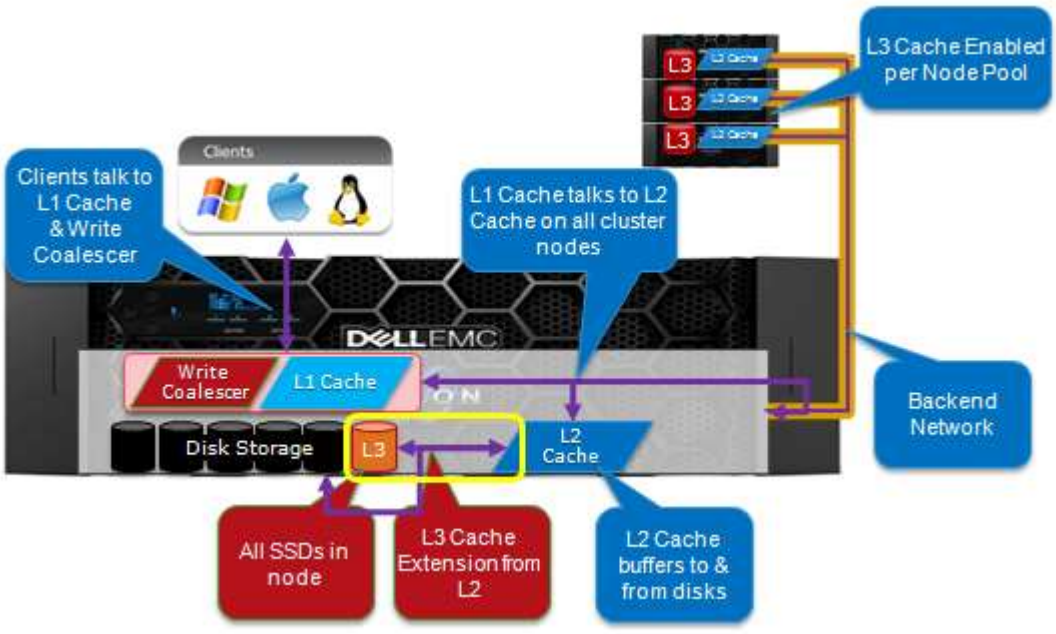


Figure 11 : Architecture de cache L1, L2 et L3 OneFS

OneFS impose qu'un fichier soit écrit sur plusieurs nœuds du cluster, et éventuellement sur plusieurs disques d'un nœud, afin que toutes les demandes de lecture impliquent la lecture de données distantes (et éventuellement locales). Lorsqu'un client envoie une demande de lecture, OneFS détermine si les données demandées sont présentes dans le cache local. Les données présentes dans le cache local sont lues immédiatement. Si les données demandées ne se trouvent pas dans le cache local, elles sont lues à partir du disque. Pour les données qui ne se trouvent pas sur le nœud local, une demande est effectuée à partir des nœuds distants sur lesquels elles résident. Une autre recherche dans le cache est effectuée sur chacun des autres nœuds. Toutes les données contenues dans le cache sont renvoyées immédiatement, et les données qui ne se trouvent pas dans le cache sont récupérées à partir du disque.

Une fois que les données ont été récupérées à partir des caches locaux et distants (et éventuellement des disques), elles sont renvoyées au client.

Les étapes générales permettant de satisfaire une demande de lecture sur les nœuds locaux et distants sont les suivantes :

Sur un nœud local (le nœud qui reçoit la demande) :

1. Déterminez si une partie des données demandées se trouvent dans le cache L1 local. Si c'est le cas, renvoyez-les au client.
2. Si elles ne se trouvent pas dans le cache local, interrogez les données à partir du ou des nœuds distants.

   Sur les nœuds distants :

1. Déterminez si les données demandées se trouvent dans le cache L2 ou L3 local. Si c'est le cas, renvoyez-les au nœud responsable de la demande.
2. Si elles ne se trouvent pas dans le cache local, lisez-les à partir du disque et renvoyez-les au nœud responsable de la demande.

La mise en cache des écritures accélère le processus d'écriture des données dans un cluster. Pour ce faire, il faut regrouper par lots les demandes d'écriture plus petites et les envoyer au disque sous forme de fragments plus importants, ce qui permet d'éliminer une quantité importante de latence d'écriture sur le disque. Lorsque les clients écrivent sur le cluster, OneFS écrit temporairement les données dans un cache de journal de la NVRAM sur le nœud initiateur, au lieu de les écrire immédiatement sur le disque. OneFS peut ensuite vider ces écritures mises en cache sur le disque plus tard, à un moment plus adéquat. De plus, ces écritures sont également mises en miroir dans les journaux NVRAM des nœuds participants pour répondre aux exigences de protection du fichier. Par conséquent, en cas de division du cluster ou de l'arrêt non planifié d'un nœud, les écritures du cache non validées sont entièrement protégées.

Le cache d'écriture fonctionne comme suit :

- Un client NFS envoie au nœud 1 la demande d'écriture d'un fichier possédant une protection +2n.
- Le nœud 1 accepte les écritures dans le cache d'écriture NVRAM (raccourci), puis les met en miroir dans les fichiers logs des nœuds participants à des fins de protection.
- Des accusés de réception d'écriture sont envoyés au client NFS immédiatement et, de ce fait, il n'y a pas de latence d'écriture sur disque.
- À mesure que le cache d'écriture du nœud 1 se remplit, il est régulièrement vidé et les écritures sont validées sur le disque par le biais du processus commit en deux phases (décrit précédemment), avec la protection par code d'effacement (ECC) appropriée appliquée (+2n).
- Les fichiers logs des nœuds participants et le cache d'écriture sont effacés et disponibles pour accepter de nouvelles écritures.

📖 Pour plus d'informations, reportez-vous au livre blanc OneFS SmartFlash.

## Lectures de fichier

Les données, les métadonnées et les inodes sont tous distribués sur plusieurs nœuds dans un cluster, voire plusieurs disques dans les nœuds. Lors de la lecture ou de l'écriture dans le cluster, le nœud auquel un client se connecte agit en tant que « capitaine » pour l'opération.

Au cours d'une opération de lecture, le nœud « capitaine » rassemble toutes les données provenant des différents nœuds du cluster et les présente de manière cohérente au demandeur.

Avec l'utilisation de matériel standard économique, le cluster fournit un rapport élevé de cache sur disque (plusieurs Go par nœud), avec allocation dynamique pour les opérations de lecture et d'écriture en fonction des besoins. Ce cache RAM est unifié et cohérent sur tous les nœuds du cluster, permettant à une demande de lecture client sur un nœud de bénéficier des E/S déjà traitées sur un autre nœud. Ces blocs mis en cache peuvent être rapidement accessibles à partir de n'importe quel nœud du fond de panier à faible latence, ce qui garantit un cache RAM de grande taille et efficace, qui accélère considérablement les performances de lecture.

Plus le cluster devient grand, plus le cache présente des avantages. C'est pour cette raison que le nombre d'E/S de disque sur un cluster est généralement nettement inférieur à celui des plates-formes traditionnelles, ce qui génère des latences réduites et une meilleure expérience utilisateur.

Pour les fichiers marqués par un schéma d'accès simultané ou en temps réel, OneFS tire parti de la lecture préalable des données basée sur l'heuristique utilisée par le composant SmartRead. SmartRead peut créer un pipeline de données à partir du cache N2, en effectuant une lecture préalable dans un cache N1 local sur le nœud « capitaine ». Cela permet d'améliorer considérablement les performances de lecture séquentielle sur tous les protocoles et signifie que les lectures proviennent directement de la RAM en quelques millisecondes. Pour les lectures hautement séquentielles, SmartRead peut exécuter une lecture préalable très agressive qui autorise les lectures et les écritures de fichiers individuels à des taux de transfert très élevés.

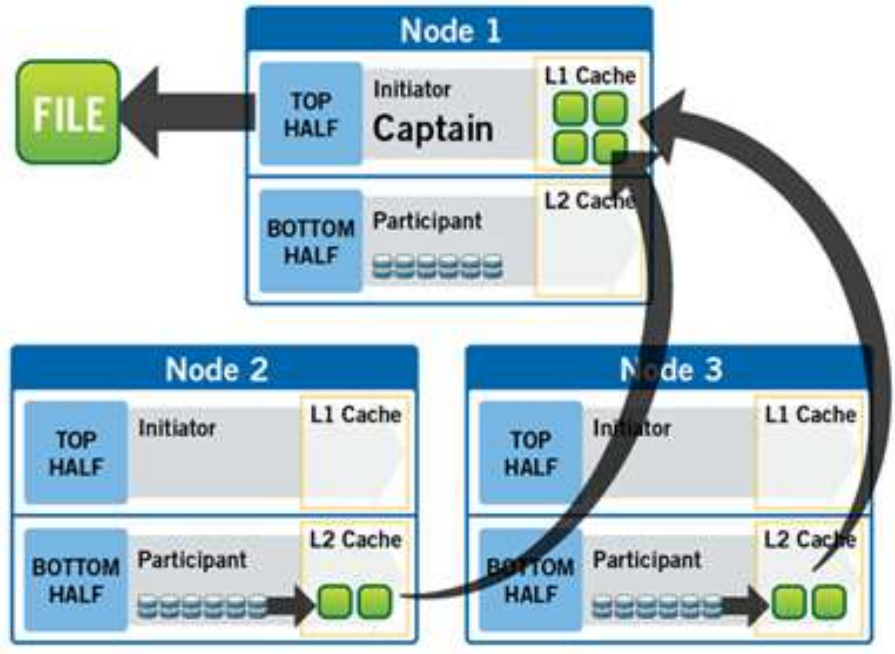


Figure 12 : Opération de lecture de fichier sur un cluster à trois nœuds

La Figure 10 montre comment SmartRead lit un fichier à accès séquentiel et non mis en cache, demandé par un client rattaché au nœud 1 dans un cluster à 3 nœuds.

1. Le nœud 1 lit les métadonnées pour identifier l'emplacement de tous les blocs de données de fichier.
2. Le nœud 1 vérifie également son cache N1 pour voir s'il contient les données de fichier demandées.
3. Le nœud 1 construit un pipeline de lecture, envoyant des demandes simultanées à tous les nœuds ayant un élément de données afin de récupérer ces données de fichier à partir du disque.
4. Chaque nœud récupère les blocs de données à partir du disque et les place dans son cache L2 (ou dans le cache SmartFlash L3, le cas échéant), puis transmet les données de fichier au nœud 1.
5. Le nœud 1 stocke les données entrantes dans le cache N1, tout en présentant simultanément le fichier au client. Dans le même temps, le processus de lecture préalable se poursuit.
6. Pour les cas hautement séquentiels, les données du cache N1 peuvent éventuellement être « abandonnées » pour libérer de la mémoire RAM pour d'autres demandes de cache N1 ou N2.

La mise en cache intelligente de SmartRead génère des performances de lecture très élevées avec des niveaux élevés d'accès simultanés. Plus important encore, il est plus facile, pour le nœud 1, d'obtenir des données de fichier du cache du nœud 2 (par le biais de l'interconnexion de cluster de faible latence) que d'accéder à son propre disque local. Les algorithmes de SmartRead contrôlent le niveau d'agressivité de la lecture préalable (lecture préalable désactivée dans les cas d'accès aléatoire) et la durée pendant laquelle les données restent dans le cache, et optimisent la mise en cache des données.

## Verrous et accès simultané

OneFS a un gestionnaire de verrous totalement distribué qui collecte les verrous des données sur l'ensemble des nœuds d'un cluster de stockage. Le gestionnaire de verrouillage est extrêmement évolutif, et permet à plusieurs « personnes » de prendre en charge les verrous du système de fichiers, ainsi que les verrous de protocole cohérents avec le cluster, tels que les verrous en mode de partage SMB ou les verrous en mode conseil NFS. OneFS permet également la prise en charge des verrous délégués tels que les oplocks CIFS et les délégations NFSv4.

Chaque nœud d'un cluster est un coordinateur de verrouillage des ressources ; un coordinateur est affecté aux ressources qu'il est possible de verrouiller basées sur un algorithme de hachage avancé. L'algorithme est conçu de telle manière que le coordinateur termine presque toujours sur un nœud différent de l'initiateur de la demande. Lorsqu'un verrou est demandé pour un fichier, il peut s'agir d'un verrou partagé (ce qui permet à plusieurs utilisateurs de partager le verrou simultanément, généralement pour les lectures) ou un verrou exclusif (pour un utilisateur, à tout moment, en général pour les écritures).

La Figure 13 ci-dessous illustre la façon dont les threads de différents nœuds peuvent demander un verrou au coordinateur.

1. Le nœud 2 est désigné coordinateur de ces ressources.
2. Le thread 1 du nœud 4 et le thread 2 du nœud 3 demandent un verrou partagé sur un fichier à partir du nœud 2, en même temps.
3. Le nœud 2 vérifie si un verrou exclusif existe pour le fichier demandé.
4. Si aucun verrou exclusif n'existe, le nœud 2 accorde au thread 1 du nœud 4 et au thread 2 du nœud 3 des verrous partagés sur le fichier demandé.
5. Les nœuds 3 et 4 exécutent maintenant une lecture sur le fichier demandé.
6. Le thread 3 du nœud 1 demande un verrou exclusif pour le fichier en cours de lecture par les nœuds 3 et 4.
7. Le nœud 2 vérifie avec les nœuds 3 et 4 si les verrous partagés peuvent être récupérés.
8. Les nœuds 3 et 4 procèdent toujours à la lecture, c'est pourquoi le nœud 2 demande au thread 3 du nœud 1 de patienter.
9. Le thread 3 du nœud 1 est bloqué jusqu'à ce qu'un verrou exclusif soit accordé par le nœud 2, puis termine l'opération d'écriture.

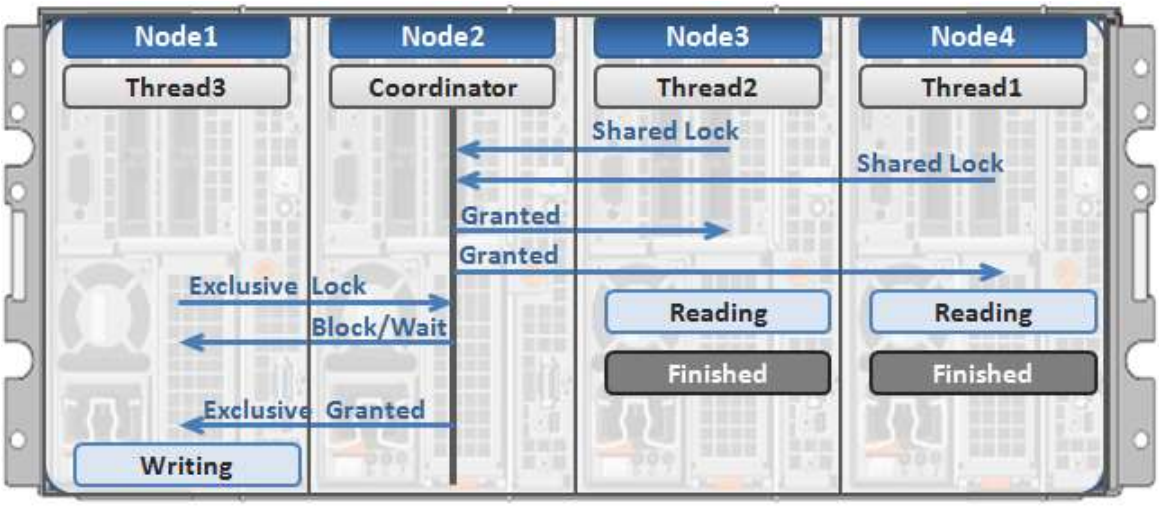


Figure 13 : Gestionnaire de verrouillage distribué

### E/S multithread

L'utilisation croissante de vastes datastores NFS pour la prise en charge de la virtualisation des serveurs et des applications d'entreprise nécessite des débits élevés et une faible latence pour les fichiers volumineux. Pour prendre en compte cette nécessité, OneFS Multi-writer prend en charge plusieurs threads capables d'écrire simultanément dans des fichiers individuels.

Dans l'exemple ci-dessus, l'accès en écriture simultané à un fichier volumineux peut être limité par le mécanisme de verrouillage exclusif, appliqué à l'ensemble du fichier. Afin d'éviter ce goulot d'étranglement potentiel, OneFS Multi-writer fournit un verrouillage en écriture plus granulaire en divisant le fichier en zones distinctes et en accordant des verrous en écriture exclusifs à des régions individuelles, et non à l'intégralité du fichier. De ce fait, plusieurs clients peuvent simultanément écrire sur différentes parties du même fichier.

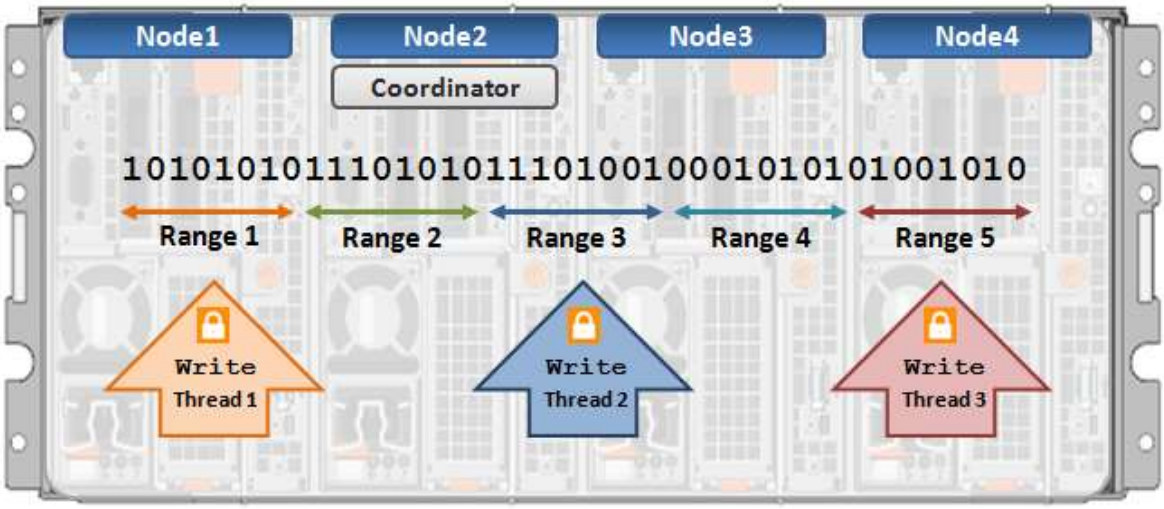


Figure 14 : Writer d'E/S multithread

# Protection des données

**Perte d'alimentation**

Un journal de système de fichiers, qui stocke des informations sur les modifications apportées au système de fichiers, est conçu pour permettre des restaurations rapides et cohérentes après des défaillances système ou des pannes, comme une perte d'alimentation. Le système de fichiers relit les entrées du journal après la restauration d'un nœud ou d'un cluster à la suite d'une panne de courant ou d'une autre panne. Sans journal, un système de fichiers devrait passer en revue individuellement toutes les modifications potentielles après une panne (opération fsck ou chkdsk). Dans un système de fichiers volumineux, cette opération peut prendre un certain temps.

OneFS est un système de fichiers basé sur la consignation dans lequel chaque nœud contient une carte NVRAM avec batterie de secours utilisée pour protéger les écritures non validées vers le système de fichiers. La batterie de la carte NVRAM dure plusieurs jours sans nécessiter de rechargement. Lorsqu'un nœud démarre, il vérifie son journal et relit de manière sélective des transactions sur le disque sur lequel le système de consignation le juge nécessaire.

OneFS effectue uniquement le montage s'il peut garantir que toutes les transactions qui ne sont pas déjà dans le système ont été enregistrées. Par exemple, si les procédures d'arrêt appropriées n'ont pas été suivies et que la batterie NVRAM est déchargée, des transactions peuvent avoir été perdues. Pour empêcher tout problème potentiel, le nœud ne monte pas le système de fichiers.

**Pannes matérielles et quorum**

Pour que le cluster fonctionne correctement et accepte les écritures de données, un quorum de nœuds doit être actif et répondre. Un quorum est défini à la majorité simple : un cluster avec des nœuds doit disposer de ⌊/2⌋+1 nœuds en ligne afin d'autoriser les écritures. Par exemple, pour un cluster de sept nœuds, quatre nœuds sont nécessaires pour obtenir un quorum. Si un nœud ou un groupe de nœuds est installé et réactif, mais n'est pas membre d'un quorum, il s'exécute en lecture seule.

OneFS utilise un quorum pour empêcher les états de déconnexion (« split-brain ») qui peuvent se produire si le cluster est temporairement divisé en deux clusters. Selon la règle de quorum, l'architecture garantit qu'indépendamment du nombre de nœuds qui tombent en panne ou sont remis en ligne, si une écriture a lieu, elle peut être rendue cohérente avec toutes les écritures précédentes qui ont déjà eu lieu. Le quorum indique également le nombre de nœuds nécessaires pour un déplacement vers un niveau de protection des données déterminé. Pour un niveau de protection basé sur un code d'effacement +, le cluster doit contenir au moins 2+1 nœuds. Par exemple, au moins sept nœuds sont requis pour une configuration +3n. Cela permet une défaillance simultanée de trois nœuds tout en maintenant un quorum de quatre nœuds pour que le cluster reste entièrement opérationnel. Si un cluster passe au-dessous du quorum, le système de fichiers est automatiquement placé dans un état de protection, en lecture seule, interdisant les écritures, mais autorisant néanmoins un accès en lecture seule aux données disponibles.

**Défaillances matérielles : ajout/suppression de nœuds**

Un système appelé protocole de gestion de groupe (GMP) active la connaissance globale de l'état du cluster à tout moment, et garantit une vue cohérente sur l'ensemble du cluster de l'état de tous les autres nœuds. Si un ou plusieurs nœuds deviennent inaccessibles via l'interconnexion de cluster, le groupe est « séparé » ou retiré du cluster. Tous les nœuds sont résolus dans une nouvelle vue cohérente de leur cluster. (C'est un peu comme si le cluster était séparé en deux groupes de nœuds distincts et qu'un seul groupe détenait le quorum.) Tant qu'elles sont dans cet état de division, toutes les données du système de fichiers sont accessibles et, pour la partie gérant le quorum, modifiables. Toutes les données stockées sur le périphérique défaillant sont reconstruites à partir de la redondance stockée dans le cluster.

Si le nœud redevient accessible, une fusion ou un ajout se produit, ce qui permet de replacer le ou les nœuds dans le cluster. (Les deux groupes fusionnent en un seul.) Le nœud peut rejoindre le cluster sans être reconstruit ni reconfiguré. Ce processus est différent de celui des baies RAID matérielles, qui nécessitent des disques pour être reconstruites. AutoBalance peut de nouveau agréger certains fichiers par bande pour gagner en efficacité, si certains de leurs groupes de protection ont été écrasés et transformés en bandes plus petites pendant la séparation.

Le moteur de tâches OneFS intègre également un processus appelé Collect, qui agit comme un collecteur d'orphelins. Lorsqu'un cluster est divisé lors d'une opération d'écriture, certains blocs qui ont été alloués au fichier devront peut-être être réaffectés à la partie quorum. Cette approche rend « orphelins » les blocs alloués à la partie sans quorum. Lorsque le cluster refusionne, la tâche Collect recherche ces blocs orphelins par une analyse « marquage et nettoyage » parallélisée et les récupère sous forme d'espace disponible pour le cluster.

**Reconstruction évolutive**

OneFS ne dépend pas de la mise en œuvre RAID matérielle pour l'allocation de données ou pour la reconstruction des données après une panne. Au lieu de cela, OneFS gère la protection des données des fichiers directement, et lorsqu'une panne se produit, OneFS reconstruit les données en mode parallélisé. OneFS peut déterminer quels fichiers sont affectés par une panne dans le temps constant, en lisant les données inode d'une manière linéaire, directement depuis le disque. L'ensemble de fichiers affectés est attribué à un jeu de threads de travail qui sont répartis entre les nœuds de cluster par le moteur de tâches. Les nœuds de traitement réparent les fichiers en parallèle. Cela implique qu'à mesure que la taille du cluster augmente, le délai de reconstruction après une panne diminue. C'est un avantage considérable en termes d'efficacité car la résilience des clusters est maintenue à mesure que leur taille augmente.

**Virtual hot spare**

La plupart des systèmes de stockage traditionnels basés sur RAID ont besoin du provisionnement d'un ou de plusieurs disques de secours pour permettre une restauration indépendante des disques défaillants. Le disque de secours remplace le disque défaillant dans un ensemble RAID. Si ces disques de secours ne sont pas eux-mêmes remplacés avant que d'autres pannes se produisent, le système risque une perte de données catastrophique. OneFS permet d'éviter l'utilisation de disques de secours et emprunte simplement de l'espace libre disponible dans le système afin de récupérer après une panne ; cette technique s'appelle disque de secours virtuel. Ce faisant, il permet au cluster de s'autoréparer entièrement, sans intervention humaine. L'administrateur peut créer une réserve de disques de secours virtuels, ce qui permet au système de s'autoréparer en dépit des écritures continues des utilisateurs.

**Protection des données en mode fichier avec codage d'effacement**

Un cluster est conçu pour tolérer une ou plusieurs échecs de composants simultanés, sans que cela empêche le cluster de servir les données. Pour y parvenir, OneFS protège les fichiers au moyen de la correction d'erreur Reed-Solomon (protection N+M) ou d'un système de mise en miroir. La protection des données est appliquée au logiciel en mode fichier, ce qui permet au système de se concentrer uniquement sur la restauration des fichiers compromis par une panne, plutôt que d'avoir à vérifier et réparer un ensemble ou un volume de fichiers entier. Les métadonnées et les inodes OneFS sont toujours protégés par la mise en miroir, plutôt que par le codage Reed-Solomon, et avec au minimum le niveau de protection des données qu'ils référencent.

Comme toutes les données, les métadonnées et les informations de protection sont distribuées entre les nœuds du cluster, le cluster n'a pas besoin d'un nœud ou d'un disque de parité dédié, ni d'un appareil ou d'un ensemble d'appareils dédié pour gérer les métadonnées. Cela garantit qu'aucun nœud ne peut devenir un point unitaire de panne. Tous les nœuds sont équitablement répartis par rapport aux tâches à exécuter, offrant ainsi une symétrie et un équilibrage de la charge idéaux dans une architecture peer-to-peer.

OneFS propose plusieurs niveaux de paramètres de protection des données configurables que vous pouvez modifier à tout moment sans avoir à mettre le cluster ou le système de fichiers hors ligne.

Pour un fichier protégé par des codes d'effacement, nous supposons que chacun de ses groupes de protection est protégé à un niveau N+M/b, où N>M et M>=b. Les valeurs N et M représentent, respectivement, le nombre de disques utilisés pour les données et les codes d'effacement au sein du groupe de protection. La valeur de b correspond au nombre de bandes de données utilisées pour un groupe de protection et est décrite ci-dessous. Un exemple courant et facile à comprendre serait un cas ou b=1, ce qui implique qu'un groupe de protection incorpore N disques de données ; M disques de redondance, stockés dans des codes d'effacement ; et que le groupe de protection doit être disposé au-dessus d'une bande au sein d'un ensemble de nœuds. Cela permet aux membres M du groupe de protection de tomber en panne simultanément tout en continuant d'offrir une disponibilité des données de 100 %. Les membres du code d'effacement M sont calculés à partir des membres de données N. La figure 13 ci-dessous illustre l'exemple d'un groupe de protection 4+2 standard (N=4, M=2, b=1).

Étant donné que OneFS répartit les fichiers entre les nœuds, cela implique que les fichiers répartis au niveau N+M peuvent résister à défaillances simultanées de nœud sans perte de disponibilité. OneFS fournit par conséquent la résilience quel que soit le type de défaillance, qu'il s'agisse d'un disque, d'un nœud ou d'un composant de nœud (une carte, par exemple). En outre, la défaillance d'un nœud compte comme une seule défaillance, quel que soit le type ou le nombre de composants défaillants. Par conséquent, si cinq disques tombent en panne dans un nœud, cela représente une seule défaillance dans le cas de la protection N+M.

OneFS peut fournir de façon unique un niveau variable pour M pouvant aller jusqu'à quatre, pour une protection des défaillances quatre fois supérieure. Cette protection va bien plus loin que le niveau RAID maximal couramment utilisé à l'heure actuelle, puisque la protection contre les défaillances est doublée par rapport au niveau RAID-6. Étant donné que la fiabilité du stockage augmente de manière géométrique avec cette quantité de redondance, la protection +4n peut être bien plus fiable que le RAID matériel traditionnel. Cette protection supplémentaire signifie que des disques SATA haute capacité, de 4 To et 6 To par exemple, peuvent être ajoutés en toute confiance.

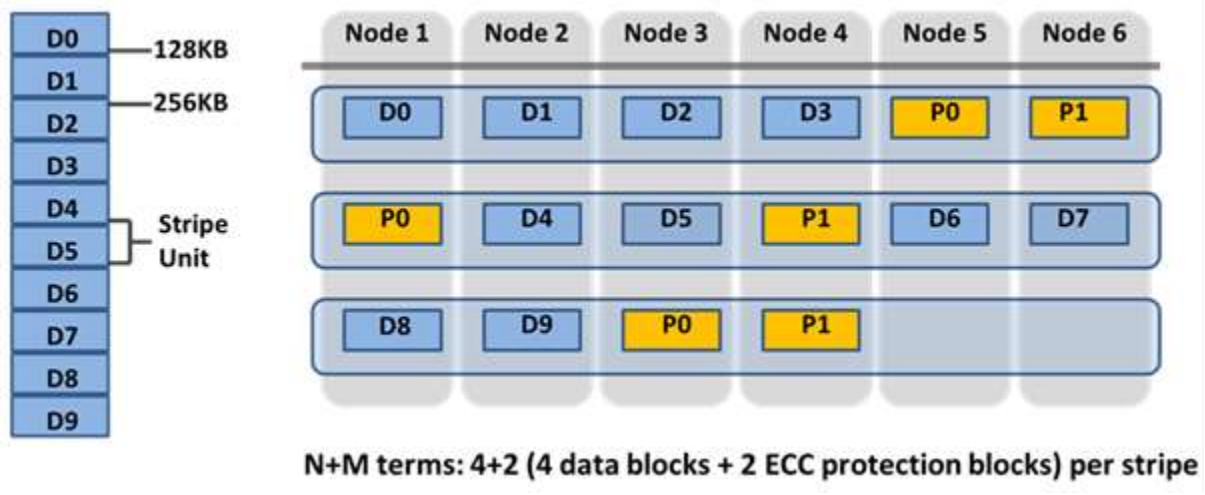


Figure 15 : Redondance OneFS : protection du code d'effacement N+M

Les clusters plus petits peuvent bénéficier d'une protection +1n, mais cela implique que si un seul disque ou nœud peut être restauré, il est impossible de restaurer deux disques présents dans deux nœuds distincts. Les défaillances de disque sont beaucoup plus susceptibles de se produire que les défaillances de nœud. Pour les clusters dotés des disques de grande taille, il est préférable d'assurer la protection contre plusieurs défaillances de disque, bien que la restauration de nœud unique soit acceptable.

Pour fournir un exemple de situation où il est souhaitable de bénéficier de la redondance de deux disques et de la redondance de nœud unique, nous pouvons créer des groupes de protection à largeur double ou triple. Ces groupes de protection à largeur double ou triple sont « appliqués » une fois ou deux au même ensemble de nœuds, lorsqu'ils sont répartis. Comme tous les groupes de protection contiennent exactement l'équivalent de deux disques de redondance, ce mécanisme permet au cluster de résister à une défaillance de deux ou trois disques ou à une défaillance du nœud complet, sans indisponibilité des données.

Le plus important pour les petits clusters, c'est que cette méthode de répartition est très efficace, avec une efficacité sur disque de M/(N+M). Par exemple, sur un cluster de cinq nœuds avec une protection contre les doubles défaillances, si nous utilisons les valeurs N=3 et M=2, nous obtenons un groupe de protection 3+2 avec une efficacité de 1−2/5, soit 60 %. Avec le même cluster à 5 nœuds mais avec chaque groupe de protection réparti sur 2 bandes, N serait à présent égal à 8 et M à 2, ce qui nous permettrait d'obtenir une efficacité sur disque de 1-2/(8+2), soit 80 %, avec maintien de la protection contre les doubles défaillances de disques, et perte uniquement de la protection contre les défaillances de deux nœuds.

OneFS prend en charge plusieurs schémas de protection, notamment le schéma +2d:1n omniprésent, qui protège contre la défaillance de deux disques ou d'un nœud.

ⓘ La pratique d'excellence consiste à utiliser le niveau de protection recommandé pour une configuration de cluster particulière. Ce niveau de protection recommandé est clairement marqué comme « suggéré » dans les pages de configuration des pools de stockage de l'interface utilisateur Web OneFS, et il est généralement configuré par défaut. Pour toutes les configurations matérielles Gen 6 actuelles, le niveau de protection recommandé est « +2d:1n ».

Les schémas de protection hybride sont particulièrement utiles pour les boîtiers Gen 6 et les configurations de nœuds haute densité, lorsque la probabilité de défaillance de plusieurs disques dépasse la probabilité de défaillance d'un nœud complet. Dans le cas peu probable où plusieurs appareils subiraient une défaillance simultanément, ce qui amènerait un fichier « au-delà de son niveau de protection », OneFS protège tous les éléments possibles et signale les erreurs sur les fichiers individuels affectés aux logs du cluster.

OneFS fournit également une large gamme d'options de mise en miroir allant de 2 à 8, ce qui permet la création de 2 à 8 miroirs du contenu spécifié. Par exemple, par défaut, les métadonnées sont mises en miroir au niveau supérieur à la correction d'erreur directe (FEC). Par exemple, si un fichier est protégé à +2n, l'objet de métadonnées qui lui est associé sera mis en miroir 3 fois.

La plage complète de niveaux de protection OneFS est résumée dans le tableau suivant :

| Niveau de protection | Description |
|---|---|
| **+1n** | Tolérer la défaillance de 1 disque OU de 1 nœud |
| **+2d:1n** | Tolère la défaillance de 2 disques OU 1 nœud |
| **+2n** | Tolère la défaillance de 2 disques OU 2 nœuds |
| **+3d:1n** | Tolère la défaillance de 3 disques OU 1 nœud |
| **+3d:1n1d** | Tolère la défaillance de 3 disques OU 1 nœud ET 1 disque |
| **+3n** | Tolère la défaillance de 3 disques OU 3 nœuds |
| **+4d:1n** | Tolère la défaillance de 4 disques OU 1 nœud |
| **+4d:2n** | Tolère la défaillance de 4 disques OU 2 nœuds |
| **+4n** | Tolère la défaillance de 4 nœuds |
| **2 à 8 fois** | Mise en miroir sur 2 à 8 nœuds, en fonction de la configuration |

OneFS permet à un administrateur de modifier la règle de protection en temps réel, alors que les clients sont connectés et lisent/écrivent des données.

ⓘ N'oubliez pas qu'augmenter le niveau de protection d'un cluster peut augmenter la quantité d'espace utilisée par les données du cluster.

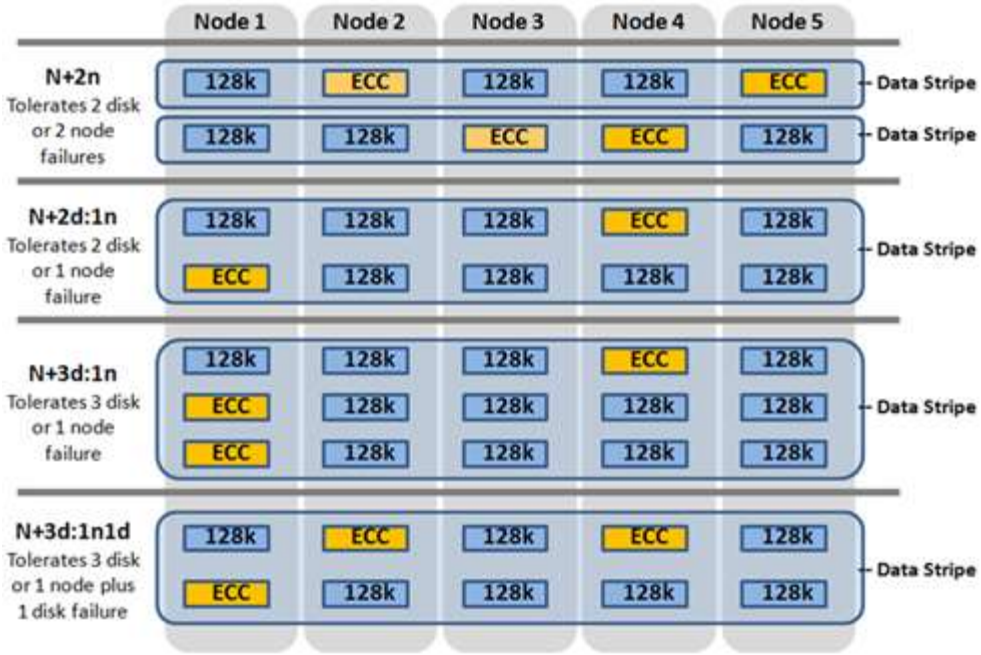


Figure 16 : Schémas de protection par code d'effacement hybride OneFS

ⓘ OneFS fournit également des alertes de sous-protection pour les nouvelles installations de cluster. Si le cluster est sous-protégé, le système de consignation des événements du cluster (CELOG) génère des alertes, avertit l'administrateur de la défaillance de protection et recommande une modification du niveau de protection approprié pour cette configuration du cluster.

🕮 Pour plus d'informations, reportez-vous au livre blanc OneFS high availability and data protection.

**Partitionnement automatique**

La hiérarchisation et la gestion des données de OneFS sont assurées par le framework SmartPools. Du point de vue de l'efficacité de la protection et de la répartition des données, SmartPools facilite la subdivision d'un nombre important de nœuds de grande capacité et homogènes en pools de disques plus petits et plus compatibles avec le ratio MTTDL (Mean Time to Data Loss, temps moyen/perte de données). Par exemple, un cluster de 80 nœuds H500 fonctionnerait généralement avec un niveau de protection +3d:1n1d. Toutefois, le partitionnement en quatre pools de disques de vingt nœuds permettrait à chaque pool de s'exécuter avec une protection +2d:1n, ce qui diminuerait le temps système de protection et améliorerait le taux d'utilisation de l'espace, sans provoquer d'augmentation nette des frais de gestion.

En accord avec l'objectif de simplicité de la gestion du stockage, OneFS calculera et partitionnera automatiquement le cluster en pools de disques, ou pools de nœuds, optimisés pour le ratio MTTDL et un taux d'utilisation efficace de l'espace. Autrement dit, les décisions relatives au niveau de protection, telles que dans l'exemple du cluster de 80 nœuds ci-dessus, ne sont pas laissées à l'appréciation du client.

Avec le provisioning automatique, chaque ensemble de matériel de nœud compatible est automatiquement divisé en pools de disques pouvant inclure jusqu'à quarante nœuds et six disques par nœud. Ces pools de nœuds bénéficient par défaut d'une protection +2d:1n. Plusieurs pools peuvent ensuite être combinés en niveaux logiques et gérés à l'aide de règles de pools de fichiers SmartPools. En subdivisant les disques d'un nœud en plusieurs pools protégés séparément, les nœuds sont plus résilients qu'auparavant aux multiples pannes de disque.

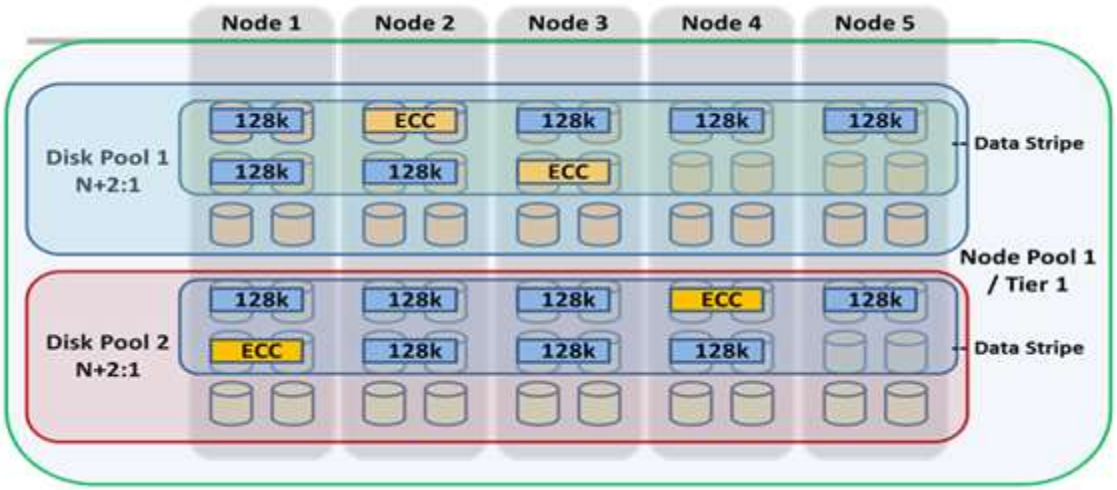


Figure 17 : Partitionnement automatique avec SmartPools

📖 Pour plus d'informations, reportez-vous au livre blanc SmartPools.

La sixième génération de plates-formes matérielles PowerScale offre une conception modulaire haute densité, où quatre nœuds sont contenus dans un seul châssis 4RU. Cette approche améliore le concept des pools de disques, des pools de nœuds et des voisinages, et elle ajoute un niveau de résilience au concept de domaine de défaillance OneFS. Chaque boîtier Gen 6 contient quatre modules de calcul (un par nœud) et cinq conteneurs de disques, ou « chariots », par nœud.

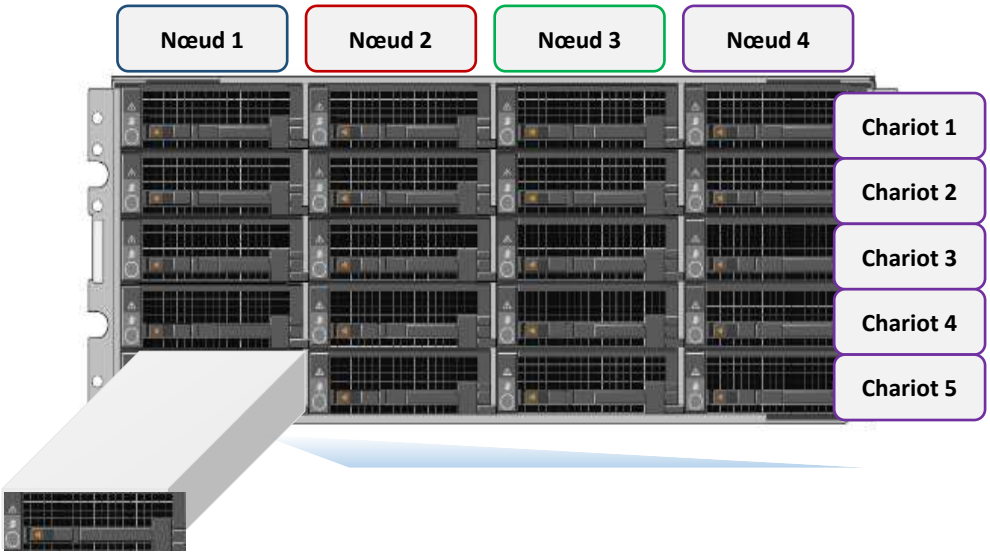


Figure 18. Vue avant du boîtier de la plate-forme Gen 6 présentant les chariots de disque.

Chaque chariot consiste en un tiroir qui se glisse à l'avant du châssis et contient entre trois et six disques, selon la configuration du châssis. Les pools de disques constituent la plus petite unité au sein de la hiérarchie des pools de stockage. Le provisioning OneFS fonctionne sur le principe de la répartition des disques de nœuds similaires en ensembles, ou en pools de disques, chaque pool représentant un domaine de défaillance distinct. Ces pools de disques sont protégés par défaut à un niveau +2d :1n (capacité de résister à la défaillance de deux disques ou d'un nœud complet).

Les pools de disques sont disposés sur les cinq chariots de chaque nœud Gen 6. Par exemple, un nœud comportant trois disques par chariot présente la configuration de pool de disques suivante :

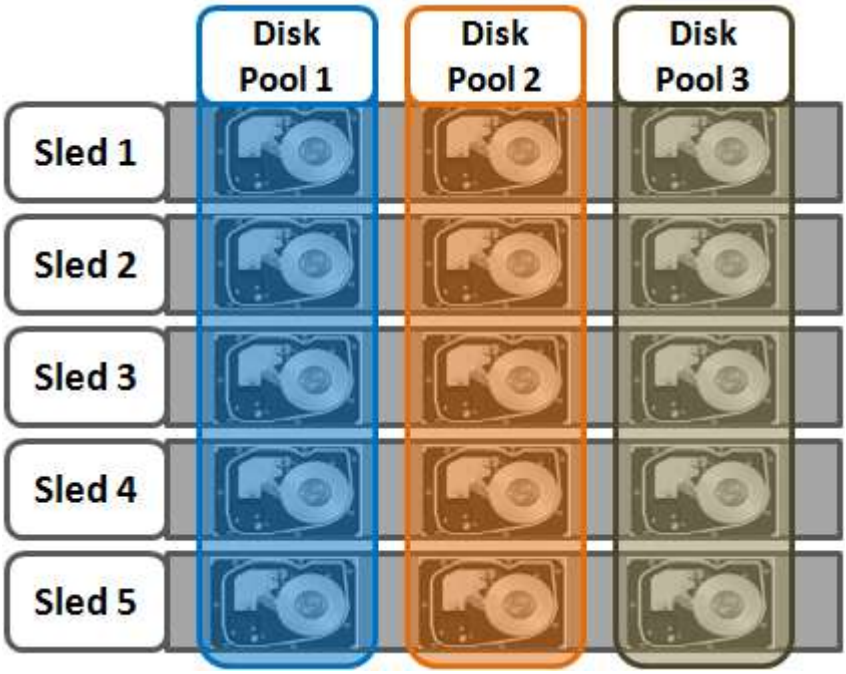


Figure 19. Pools de disques OneFS

Les pools de nœuds sont des groupes de pools de disques, répartis sur des nœuds de stockage similaires (classes de compatibilité). Cela est illustré dans la Figure 20 ci-dessous. Plusieurs groupes de différents types de nœuds peuvent fonctionner ensemble dans un cluster hétérogène unique. Par exemple : un pool de nœuds F-Series pour les applications gourmandes en opérations d'E/S, un pool de nœuds H-Series, principalement utilisé pour les charges applicatives simultanées et séquentielles, et un pool de nœuds A-Series, principalement utilisé pour les charges applicatives d'archivage nearline et/ou à long terme.

Cela permet à OneFS de présenter un seul pool de ressources de stockage comprenant plusieurs types de supports de disques (disque SSD, SAS haute vitesse, SATA grande capacité, etc.) fournissant différents niveaux de performances, de protection et de capacité. Ce pool de stockage hétérogène peut à son tour prendre en charge un large éventail d'applications et de besoins en matière de charges de travail au travers d'un point de gestion unique et unifié. Il facilite également la combinaison de matériel ancien et récent, ce qui simplifie la protection des investissements, même pour différentes générations de produits, et l'actualisation fluide du matériel.

Chaque pool de nœuds contient uniquement des pools de disques issus du même type de nœuds de stockage, et un pool de disques peut appartenir à un seul pool de nœuds. Par exemple, un pool de nœuds pourrait contenir tous les nœuds de la F-Series dotés de disques SSD de 1,6 To, tandis qu'un autre comprendrait les nœuds A-Series dotés de disques SATA de 10 To. Aujourd'hui, un minimum de 4 nœuds (un boîtier) est requis par pool de nœuds pour le matériel Gen 6, tel que le PowerScale H700, ou trois nœuds par pool pour les nœuds autonomes tels que le PowerScale F900.

Les voisinages OneFS sont des domaines de défaillance au sein d'un pool de nœuds, et leur objectif est d'améliorer la fiabilité en général et de protéger les données contre toute indisponibilité des données lors de la suppression accidentelle de chariots de disque. Pour les nœuds autonomes tels que PowerScale F200, OneFS a une taille idéale de 20 nœuds par pool de nœuds et une taille maximale de 39 nœuds. Lors de l'ajout du 40e nœud, les nœuds étaient divisés en deux voisinages de 20 nœuds.

Avec la plate-forme Gen 6, la taille idéale d'un voisinage passe de 20 à 10 nœuds. Cette caractéristique protège contre les défaillances simultanées du journal des paires de nœuds et les défaillances complètes du châssis.

Les nœuds partenaires sont des nœuds dont les journaux sont mis en miroir. Avec la plate-forme Gen 6, au lieu que chaque nœud stocke son journal dans la NVRAM comme dans les plates-formes précédentes, les journaux des nœuds sont stockés sur des disques SSD, et chaque journal dispose d'une copie miroir sur un autre nœud. Le nœud qui contient le journal en miroir est appelé nœud partenaire. Les modifications apportées au stockage des journaux présentent plusieurs avantages en matière de fiabilité. Par exemple, les disques SSD sont plus persistants et plus fiables que la NVRAM, qui a besoin d'une batterie chargée pour conserver les informations relatives à l'état. En outre, avec le journal en miroir, les deux disques de journal doivent être arrêtés avant qu'un journal ne soit considéré comme perdu. Par conséquent, à moins que les deux disques de journal en miroir ne tombent en panne, les deux nœuds partenaires peuvent fonctionner normalement.

Dans le cadre de la protection par nœud partenaire, si possible, les nœuds sont placés dans des voisinages différents, et donc des domaines de défaillance différents. La protection des nœuds partenaires est possible dès que le cluster compte cinq châssis complets (20 nœuds) lorsque, après la première division de voisinage, OneFS place les nœuds de partenariat dans des voisinages différents :

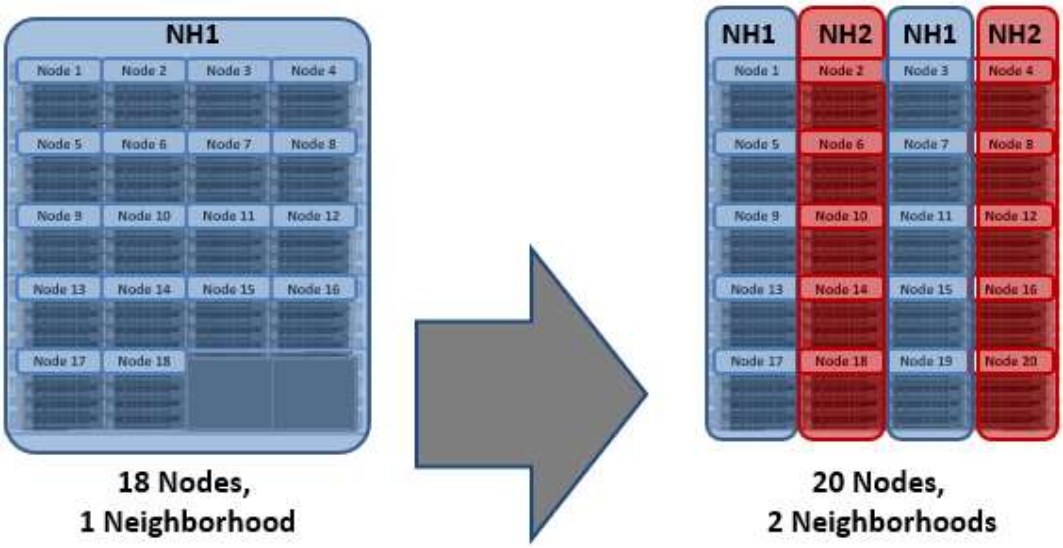


Figure 20. Division en deux voisinages de vingt nœuds.

La protection par nœuds partenaires améliore la fiabilité, car si les deux nœuds tombent en panne, ils se trouvent dans des domaines de défaillance différents, et leurs domaines de défaillance respectifs ne subissent donc que la perte d'un seul nœud.

Avec la protection du châssis, le cas échéant, chacun des quatre nœuds au sein d'un châssis sera placé dans un voisinage séparé. La protection du châssis devient possible à 40 nœuds, car la division en voisinages de 40 nœuds permet de placer chaque nœud d'un châssis dans un autre voisinage. Par conséquent, lorsqu'un cluster Gen 6 de 38 nœuds est étendu à 40 nœuds, les deux voisinages existants sont divisés en voisinages de 10 nœuds :

La protection du châssis garantit qu'en cas de défaillance du châssis, chaque domaine de défaillance ne perd qu'un seul nœud.

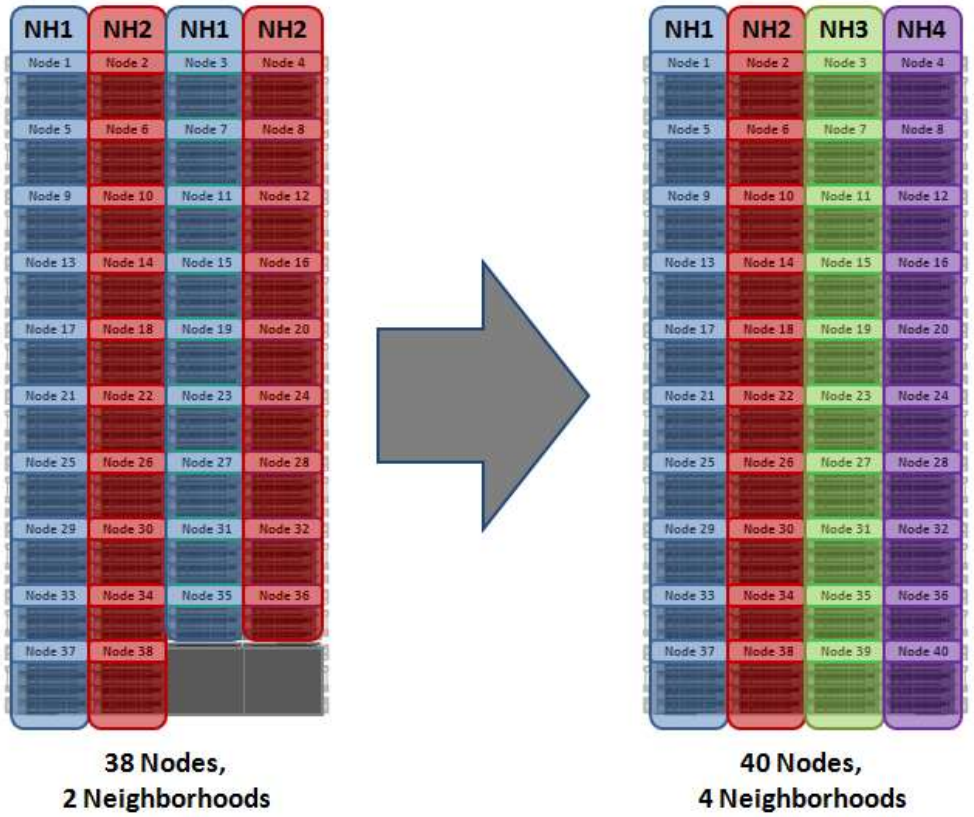


Figure 21. Voisinages OneFS : division en quatre voisinages.

ⓘ Un cluster de 40 nœuds ou plus avec 4 voisinages, protégé conformément au niveau par défaut +2d :1n, peut supporter une défaillance de nœud unique par voisinage. Cela protège le cluster en cas de défaillance d'un seul châssis Gen 6.

Globalement, un cluster de plate-forme Gen 6 a une fiabilité supérieure à celle des clusters de génération précédente et de capacité similaire, en raison des améliorations suivantes :

- Journaux en miroir
- Voisinages plus petits
- Disques de démarrage mis en miroir

## Compatibilité

Certains types de nœuds similaires, mais non identiques, peuvent être provisionnés sur un pool de nœuds existant par compatibilité de nœud. Avec OneFS, un pool de nœuds doit comporter au minimum trois nœuds.

ⓘ En raison d'importantes différences architecturales, il n'existe aucune compatibilité de nœud entre les plates-formes Gen 6, les générations matérielles précédentes ou les nœuds PowerScale.

OneFS contient également une option de compatibilité de disque SSD, qui permet de provisionner les nœuds disposant de disques SSD de capacités différentes vers un pool de nœuds unique.

La compatibilité de disque SSD est créée et décrite dans la liste de compatibilités SmartPools de l'interface utilisateur Web OneFS, et elle apparaît également dans la liste des niveaux et pools de nœuds.

ⓘ Lors de la création de cette compatibilité de disque SSD, OneFS vérifie automatiquement que les deux pools à fusionner ont les mêmes nombres de disques SSD (le cas échéant), niveaux, protection requise et paramètres de cache L3. Si ces paramètres diffèrent, l'interface utilisateur Web OneFS vous invite à consolider et aligner ces paramètres.

📖 Pour plus d'informations, reportez-vous au livre blanc SmartPools.

## Protocoles pris en charge

Les clients disposant des informations d'identification et des privilèges appropriés peuvent créer, modifier et lire les données à l'aide de l'une des méthodes standard pour communiquer avec le cluster :

- NFS (Network File System)
- SMB/CIFS (Server Message Block/Common Internet File System)
- FTP (File Transfer Protocol)
- HTTP (Hypertext Transfer Protocol)
- HDFS (Hadoop Distributed File System)
- REST API (Representational State Transfer Application Programming Interface)
- S3 (API de stockage en mode objet)

Pour le protocole NFS, OneFS prend en charge NFSv3 et NFSv4, ainsi que NFSv4.1 dans OneFS 9.3. De plus, OneFS 9.2 et les versions ultérieures incluent la prise en charge de NFSv3overRDMA.

Pour Microsoft Windows, le protocole SMB est pris en charge jusqu'à la version 3. Dans le cadre de SMB3, OneFS prend en charge les fonctions suivantes :

- Multipathing SMB3
- Disponibilité continue et Witness SMB3
- Chiffrement SMB3

Le chiffrement SMB3 peut être configuré par part ou à l'échelle du cluster ou de la zone. Seuls les systèmes d'exploitation qui prennent en charge le chiffrement SMB3 peuvent fonctionner avec des partages chiffrés. Ces systèmes d'exploitation peuvent également fonctionner avec des partages non chiffrés si le cluster est configuré de façon à autoriser les connexions non chiffrées. D'autres systèmes d'exploitation peuvent aussi accéder à des partages non chiffrés, uniquement si le cluster est configuré de façon à autoriser les connexions non chiffrées.

La racine du système de fichiers pour toutes les données du cluster est /ifs (système de fichiers OneFS). Le protocole SMB utilise la présentation de type partage « ifs » (\\<cluster_name\ifs), et le protocole NFS utilise la présentation de type « /ifs » (<cluster_name>:/ifs).

ⓘ Les données étant communes à tous les protocoles, les modifications apportées au contenu des fichiers par le biais d'un protocole d'accès sont instantanément affichables par tous les autres.

OneFS offre une prise en charge complète des environnements IPv4 et IPv6 sur les réseaux Ethernet front-end, SmartConnect et la gamme complète de protocoles de stockage et d'outils de gestion.

En outre, OneFS CloudPools prend en charge les API de stockage des fournisseurs de Cloud suivantes, ce qui permet de convertir les fichiers en plusieurs cibles de stockage, notamment :

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

📖 Pour plus d'informations, reportez-vous au guide d'administration CloudPools.

## Opérations sans perturbation : prise en charge des protocoles

OneFS participe à la disponibilité des données en prenant en charge le basculement sur incident NFSv3 et NFSv4 dynamique et le retour arrière pour les clients Linux et UNIX, ainsi que la disponibilité continue SMB3 pour les clients Windows. Cette technique garantit, en cas de défaillance de nœud ou d'opération de maintenance préventive, le transfert de toutes les opérations de lecture/écriture en cours vers un autre nœud du cluster afin de terminer l'opération sans perturber les utilisateurs ni les applications.

Lors d'un basculement sur incident, les clients sont répartis uniformément sur tous les nœuds restants du cluster, pour un impact minimal sur les performances. Si un nœud est mis hors ligne pour une raison quelconque, y compris pour une défaillance, les adresses IP virtuelles sur ce nœud sont migrées en toute transparence vers un autre nœud du cluster.

Lorsque le nœud hors ligne est remis en ligne, SmartConnect rééquilibre automatiquement les clients NFS et SMB3 sur l'ensemble du cluster afin de garantir une utilisation optimale du stockage et des performances. Dans le cadre des mises à jour logicielles et de la maintenance périodique du système, cette fonctionnalité permet de procéder à des mises à niveau consécutives par nœud, pour une disponibilité complète pendant toute la durée de la fenêtre de maintenance.

## Filtrage des fichiers

Le filtrage de fichiers OneFS peut être utilisé sur les clients NFS et SMB afin d'autoriser ou d'interdire les écritures vers une zone d'exportation, de partage ou d'accès. Cette fonction empêche le blocage de certains types de fichiers qui peuvent provoquer des problèmes de sécurité, des interruptions de la productivité, des problèmes de débit ou un encombrement de stockage. La configuration peut se faire par le biais d'une liste d'exclusion qui bloque les extensions de fichiers explicites, ou d'une liste d'inclusion qui permet explicitement les écritures de certains types de fichiers uniquement.

## Déduplication des données : SmartDedupe

Le produit SmartDedupe optimise l'efficacité du stockage d'un cluster en réduisant la capacité de stockage physique nécessaire pour héberger les données d'une organisation. De plus, l'analyse des données sur disque permet de rechercher les blocs identiques, puis d'éliminer les données en double, pour une plus grande efficacité. Cette approche est généralement appelée déduplication post-traitement, ou asynchrone.

Une fois les blocs en double découverts, SmartDedupe déplace une seule copie de ces blocs vers un ensemble de fichiers spécifiques appelé « zone de stockage de clichés instantanés ». Au cours de ce processus, les blocs en double sont supprimés des fichiers et remplacés par des pointeurs dirigés vers les zones de stockage de clichés instantanés.

Avec la déduplication post-traitement, les nouvelles données sont d'abord stockées sur le périphérique de stockage, avec d'être soumises à un processus d'analyse visant à repérer les redondances. Cela signifie que les performances initiales d'écriture ou de modification des fichiers ne sont pas affectées, car aucun calcul supplémentaire n'est nécessaire dans le chemin d'écriture.

#### Architecture SmartDedupe

L'architecture OneFS SmartDedupe se compose de cinq modules principaux :

- Control path de la déduplication
- Tâche de déduplication
- Moteur de déduplication
- Zone de stockage de clichés instantanés
- Infrastructure de déduplication

Le Control Path SmartDedupe comprend l'interface de gestion Web (WebUI) OneFS, l'interface de ligne de commande (CLI) et l'API de plate-forme RESTful, et il est responsable de la gestion de la configuration, de la planification et du contrôle de la tâche de déduplication. Il s'agit d'un processus d'arrière-plan hautement distribué qui gère l'orchestration de la déduplication sur tous les nœuds du cluster. Le contrôle des tâches englobe l'analyse du système de fichiers, et la détection et le partage des blocs de données correspondants, en collaboration avec le moteur de déduplication. La couche d'infrastructure de déduplication est le module de noyau qui effectue la consolidation des blocs de données partagées dans les zones de stockage de clichés instantanés, c'est-à-dire les conteneurs du système de fichiers qui contiennent à la fois des blocs de données physiques et des références, ou des pointeurs, vers des blocs partagés. Ces éléments sont décrits plus en détail ci-dessous.

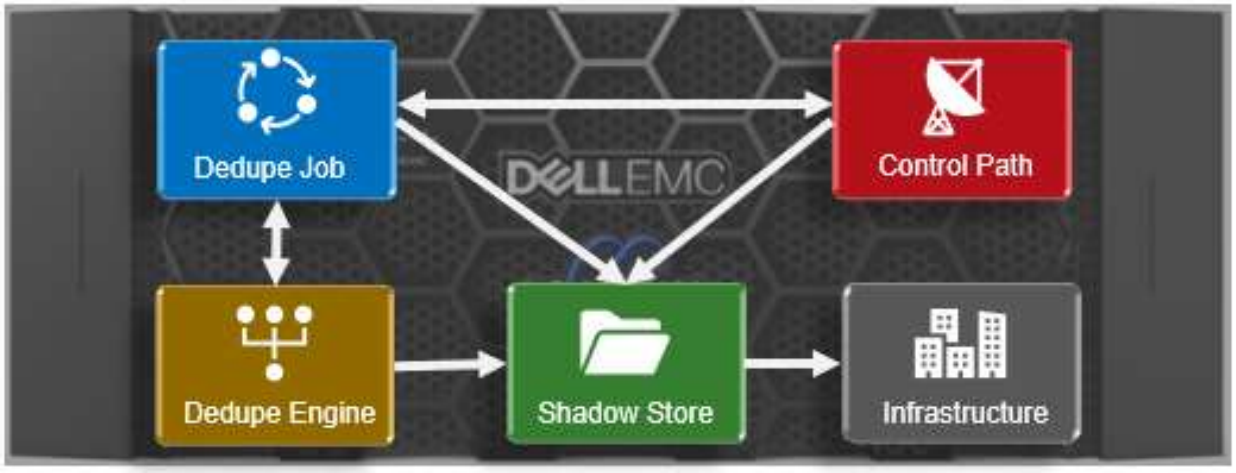


Figure 22 : Architecture modulaire OneFS SmartDedupe

📖 Pour plus d'informations, reportez-vous au livre blanc OneFS SmartDedupe.

**Zones de stockage de clichés instantanés**

Les zones de stockage de clichés instantanés OneFS sont des conteneurs de système de fichiers qui permettent de stocker les données de manière partagée. Par conséquent, les fichiers sur OneFS peuvent contenir à la fois des données physiques et des pointeurs, ou des références, vers des blocs partagés dans des zones de stockage de clichés instantanés.

Les zones de stockage de clichés instantanés sont semblables aux fichiers standard, mais elles ne contiennent pas toutes les métadonnées généralement associées à des inodes de fichiers standard. En particulier, les attributs basés sur le temps (heure de création, heure de modification, etc.) ne sont pas conservés explicitement. Chaque zone de stockage de clichés instantanés peut contenir jusqu'à 256 blocs, chaque bloc pouvant être référencé par 32 000 fichiers. Si cette limite de référence de 32 000 fichiers est dépassée, une zone de stockage de clichés instantanés est créée. En outre, les zones de stockage ne font pas référence à d'autres zones de stockage. Enfin, les snapshots de zones de stockage de clichés instantanés ne sont pas autorisés, étant donné que celles-ci n'ont pas de liens matériels.

ⓘ Les magasins d'instantanés sont également utilisés pour les fichiers clones OneFS et l'efficacité du stockage des fichiers de petite taille (SFSE), ainsi que la déduplication.

## Efficacité du stockage des fichiers de petite taille

L'efficacité du stockage des fichiers de petite taille OneFS est un autre principe consommateur de zones de stockage de clichés instantanés. Cette fonctionnalité optimise l'utilisation de l'espace d'un cluster en diminuant la quantité de stockage physique requise pour héberger les petits fichiers qui composent un jeu de donnée archivé, comme ceux des workflows PACS dans le domaine des services de santé.

L'efficacité est assurée par l'analyse des données sur disque pour détecter les petits fichiers qui sont protégés par des miroirs de copie complète et leur compression dans des zones de stockage de clichés instantanés. Ces zones de stockage de clichés instantanés sont ensuite protégées par parité et non en miroir, et offrent généralement une efficacité de stockage de 80 % ou plus.

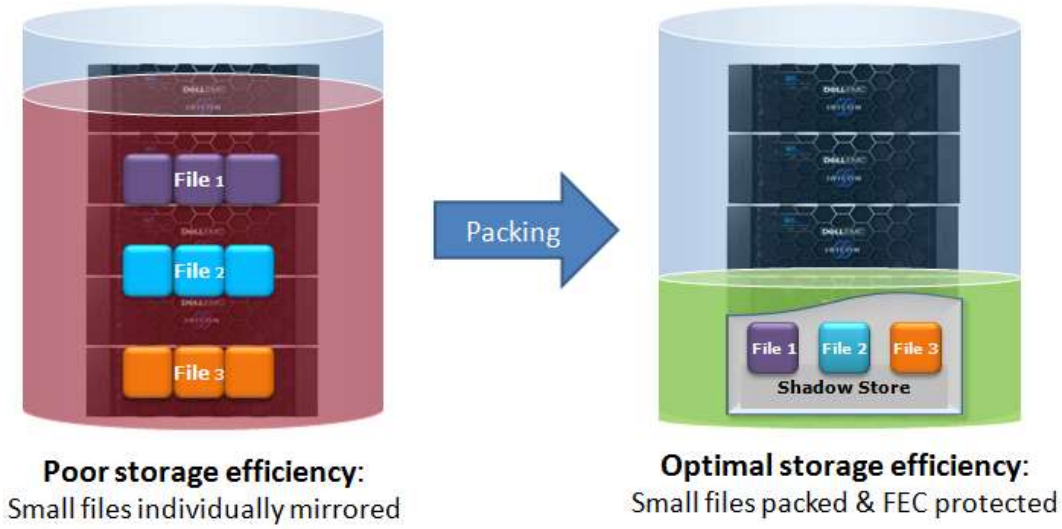


Figure 23 : Conteneurisation de fichiers de petite taille

L'efficacité du stockage des fichiers de petite taille fait baisser les performances en matière de latence de lecture des fichiers de petite taille afin d'améliorer le taux d'utilisation du stockage. Les fichiers archivés restent évidemment accessibles en écriture, mais lorsque des fichiers de conteneur qui incluent des références à des stockages de clichés instantanés sont supprimés, tronqués ou remplacés, ils peuvent laisser des blocs non référencés dans les zones de stockage de clichés instantanés. Ces blocs sont libérés ultérieurement et peuvent générer des trous qui réduisent l'efficacité du stockage.

La perte d'efficacité réelle dépend de la répartition du niveau de protection utilisée par la zone de stockage de clichés instantanés. Les petits groupes de protection sont plus sensibles, tout comme les fichiers en conteneurs, car tous les blocs des conteneurs ont au maximum un fichier de référence, et les tailles compressées (des fichiers) sont minimes.

Un défragmenteur réduit la fragmentation des fichiers due aux remplacements et aux suppressions. Ce défragmenteur de zones de stockage de clichés instantanés est intégré dans la tâche ShadowStoreDelete. Le processus de défragmentation divise chaque fichier de conteneur en fragments logiques (d'environ 32 Mo chacun) et évalue chaque fragment pour la fragmentation.

Si l'efficacité du stockage d'un fragment fragmenté est inférieure à la cible, les données sont évacuées vers un autre emplacement. La valeur cible d'efficacité par défaut est de 90 % de l'efficacité maximale du stockage disponible avec le niveau de protection utilisé par le shadow store. Des groupes de protection plus grands peuvent tolérer un niveau de fragmentation plus élevé avant que l'efficacité du stockage ne tombe en dessous de ce seuil.

## Réduction des données à la volée

La réduction des données à la volée OneFS est disponible sur les nœuds All-Flash F900, F810, F600 et F200, les boîtiers hybrides H700/7000 et H5600, ainsi que sur la plate-forme d'archivage A300/3000. L'architecture OneFS comporte les composants principaux suivants :

- Plate-forme de réduction des données
- Moteur de compression et mappage des fragments
- Phase de suppression du bloc zéro
- Index de déduplication en mémoire et infrastructure de stockage des clichés instantanés
- Framework d'alerte et de création de rapport pour la réduction des données
- Chemin de contrôle de réduction des données

Le chemin d'écriture de réduction des données à la volée se compose de trois phases principales :

- Suppression du bloc zéro
- Déduplication à la volée
- Compression à la volée

Si la compression et la déduplication à la volée sont activées sur un cluster, la suppression du bloc zéro est exécutée en premier, suivie de la déduplication, puis de la compression. Cet ordre permet à chaque phase de réduire le périmètre de travail de chaque phase suivante.


Figure 24 : Workflow de réduction des données à la volée.

Le modèle F810 inclut une fonctionnalité de déchargement de compression matérielle, et chaque nœud du châssis F810 contient un adaptateur Mellanox Innova-2 Flex. Autrement dit, la compression et la décompression sont effectuées de manière transparente par l'adaptateur Mellanox avec des temps de latence minimaux, ce qui évite d'utiliser les ressources onéreuses de processeur et de mémoire d'un nœud.

Le système de compression matérielle OneFS utilise zlib, avec une implémentation logicielle igzip pour PowerScale F900, F810, F600, F200, H700/7000, H5600, et les nœuds A300/3000. La compression logicielle est également utilisée comme solution de secours en cas de panne de la compression matérielle. Dans un cluster mixte, elle est utilisée dans des nœuds non F810 sans fonctionnalité de compression matérielle, et comme une solution de secours en cas de panne de la compression matérielle. OneFS utilise des fragments de compression de 128 Ko, et chaque fragment contient 16 blocs de données de 8 Ko. Cette méthode est optimale, car la taille des fragments est identique à celle utilisée par OneFS pour ses unités de bande de protection des données, ce qui garantit simplicité et efficacité, tout en évitant le temps système de la compression supplémentaire des fragments.

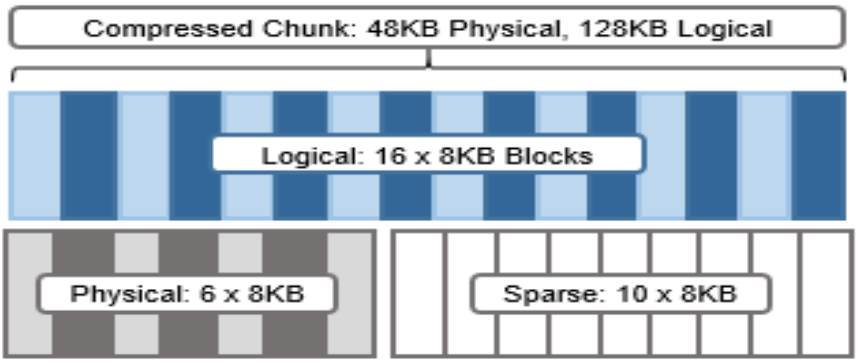


Figure 25 : Fragments de compression et superposition transparente OneFS.

Examinez le schéma ci-dessus. Après la compression, la taille de ce fragment passe de 16 à 6 blocs de 8 Ko. Autrement dit, ce fragment a maintenant une taille physique de 48 Ko. OneFS fournit une superposition logique transparente pour les attributs physiques. Cette superposition indique si les données de sauvegarde sont compressées ou non, et quels blocs du fragment sont physiques ou éparpillés, de sorte que les utilisateurs du système de fichiers ne sont pas touchés par la compression. Par conséquent, le fragment compressé est représenté de manière logique comme ayant une taille de 128 Ko, quelle que soit sa taille physique réelle.

Les gains d'efficacité doivent être au moins de 8 Ko (un bloc) pour que la compression s'effectue. Sinon, le fragment ou le fichier n'est pas pris en compte et il reste dans son état d'origine non compressé. Par exemple, un fichier de 16 Ko qui génère 8 Ko (un bloc) d'économies sera compressé. Une fois qu'un fichier a été compressé, il bénéficie ensuite d'une protection FEC (correction d'erreur directe).

Les fragments de compression ne passent jamais d'un pool de nœuds à un autre. Il n'est donc plus nécessaire de décompresser ou de recompresser les données pour modifier les niveaux de protection, d'effectuer des écritures restaurées ou de modifier les limites du groupe de protection.

**Évolutivité dynamique/Scale On Demand**

**Performances et capacité**

Contrairement aux systèmes de stockage traditionnels qui doivent évoluer verticalement (scale-up) lorsque les performances ou la capacité atteignent leurs limites, OneFS permet à un système de stockage d'évoluer horizontalement (scale-out), en élargissant de manière transparente le système de fichiers ou le volume existant pour atteindre une capacité de plusieurs pétaoctets, tout en augmentant parallèlement les performances de façon linéaire.

L'augmentation de la capacité et des performances est beaucoup plus facile sur un cluster que sur les autres systèmes de stockage. Elle ne demande que trois étapes simples à l'administrateur de stockage : ajouter un autre nœud au rack, connecter le nœud au réseau et demander au cluster d'ajouter le nœud supplémentaire. Le nouveau nœud offre de la capacité et des performances supplémentaires, puisque chaque nœud intègre des ressources CPU, mémoire, cache, réseau, NVRAM et des chemins de contrôle des E/S.

La fonction AutoBalance de OneFS déplace les données sur le réseau de manière automatique et cohérente, de telle sorte que les données existantes résidant sur le cluster sont stockées sur ce nouveau nœud de stockage. Ce rééquilibrage automatique évite que ce nœud ne devienne un point sensible pour les nouvelles données et permet aux données existantes de bénéficier des avantages d'un système de stockage plus puissant. La fonction AutoBalance de OneFS est également totalement transparente pour l'utilisateur. Il est possible de la régler pour minimiser son impact sur les charges de travail hautes performances. À elle seule, cette fonctionnalité permet à OneFS de passer, de manière transparente et à la volée, d'une capacité de téraoctets à téraoctets, sans augmenter le temps de gestion de l'administrateur ni la complexité du système de stockage.

Un système de stockage à grande échelle doit offrir les performances requises pour différents types de workflows, qu'ils soient séquentiels, simultanés ou aléatoires. Il existe différents workflows entre les applications et au sein d'une application. À travers un logiciel intelligent, OneFS répond à tous ces besoins à la fois. Plus important encore, avec OneFS, le débit et le nombre d'E/S par seconde évoluent de façon linéaire en fonction du nombre de nœuds présents dans le système. Grâce à la distribution équilibrée des données, au rééquilibrage automatique et au traitement distribué, OneFS est en mesure de tirer le meilleur parti des CPU, des ports réseau et de la mémoire supplémentaires à mesure que le système évolue.

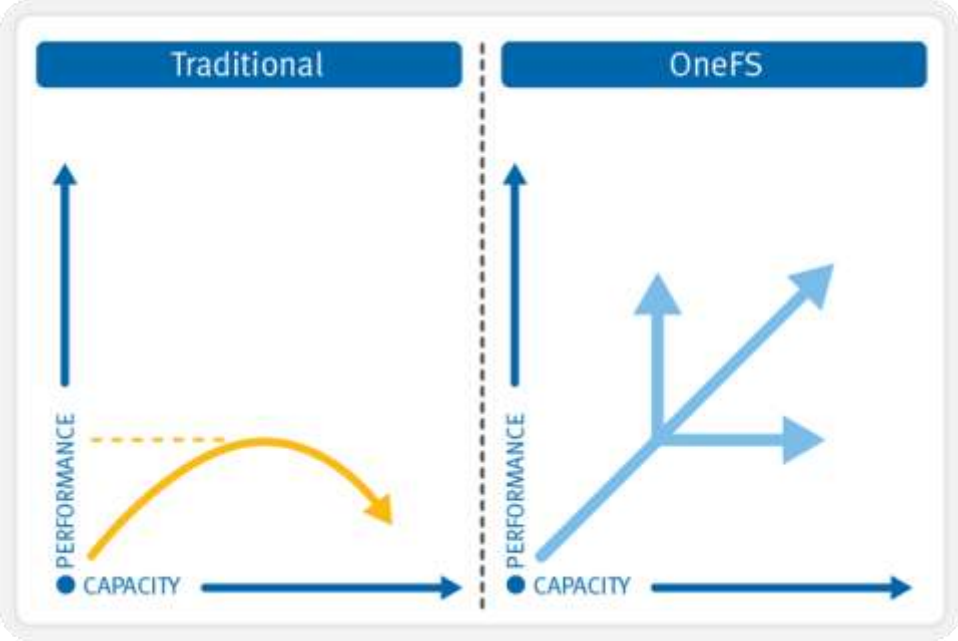


Figure 26 : Évolutivité linéaire OneFS

## Interfaces

Les administrateurs peuvent utiliser plusieurs interfaces pour gérer un cluster de stockage dans leur environnement :

- Interface utilisateur d'administration Web (WebUI)

- Interface de ligne de commande via l'accès réseau SSH ou la connexion série RS232
- Écran LCD sur les nœuds eux-mêmes pour des fonctions d'ajout/de suppression simples
- API RESTful de plate-forme pour le contrôle et l'automatisation par programmation de la configuration et de la gestion du cluster.

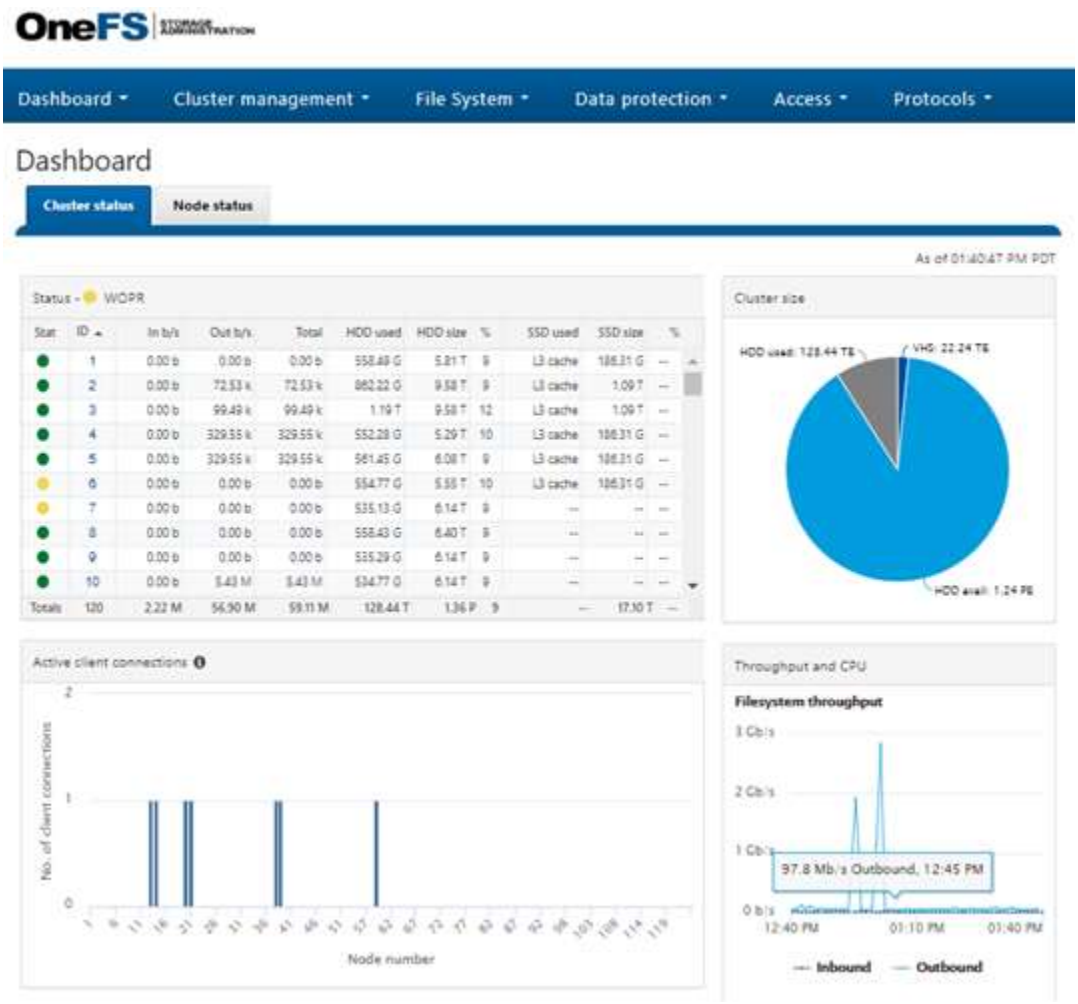


Figure 27 : Interface utilisateur Web OneFS

📖 Pour plus d'informations sur la configuration des commandes et des fonctionnalités de OneFS, reportez-vous au guide d'administration OneFS.

## Authentification et contrôle d'accès

Les services d'authentification offrent une couche de sécurité en vérifiant les informations d'identification des utilisateurs avant de leur permettre de consulter et modifier des fichiers. OneFS prend en charge quatre méthodes d'authentification des utilisateurs :

- Active Directory (AD)
- LDAP (Lightweight Directory Access Protocol)
- NIS (Network Information Service)
- Groupes et utilisateurs locaux

OneFS permet d'utiliser plusieurs types d'authentification. Toutefois, nous vous recommandons de vous familiariser en détail avec les interactions qui existent entre ces types d'authentification avant d'activer plusieurs méthodes sur le cluster. Consultez la documentation du produit pour savoir comment configurer correctement plusieurs modes d'authentification.

## Active Directory

Active Directory, une implémentation Microsoft de LDAP, est un service d'annuaire qui peut stocker des informations relatives aux ressources réseau. Bien qu'Active Directory puisse jouer de nombreux rôles, l'intérêt principal de l'association du cluster au domaine réside dans la mise en œuvre de l'authentification des utilisateurs et des groupes.

Vous pouvez configurer et gérer les paramètres Active Directory d'un cluster à partir de l'interface d'administration Web ou de l'interface CLI ; toutefois, il est généralement recommandé d'utiliser l'interface d'administration Web pour cette opération.

Chaque nœud du cluster partage le même compte de machine Active Directory, ce qui facilite considérablement l'administration et la gestion.

### LDAP

Le protocole LDAP (Lightweight Directory Access Protocol) est un protocole réseau utilisé pour définir, interroger et modifier des services et des ressources. Son principal avantage réside dans le caractère ouvert des services d'annuaire et dans la possibilité de l'utiliser sur un grand nombre de plates-formes. Le système de stockage en cluster peut utiliser LDAP pour authentifier les utilisateurs et les groupes en vue de les autoriser à accéder au cluster.

### NIS

NIS, ou Network Information Service, est un protocole de services d'annuaire conçu par Sun Microsystems. Il peut être utilisé par OneFS pour authentifier les utilisateurs et les groupes lors de l'accès au cluster. NIS, parfois appelé Yellow Pages (YP), est différent de NIS+, que OneFS ne prend pas en charge.

### Utilisateurs locaux

OneFS prend en charge l'authentification des utilisateurs locaux et des groupes. Vous pouvez créer des comptes utilisateur et de groupe locaux directement sur le cluster, à l'aide de l'interface WebUI. L'authentification locale peut se révéler utile si vous n'utilisez pas de services d'annuaire tels qu'Active Directory, LDAP ou NIS, ou lorsqu'un utilisateur ou une application spécifique a besoin d'accéder au cluster.

## Access Zones

Les zones d'accès permettent de partitionner logiquement l'accès au cluster et d'allouer les ressources pour les unités autonomes, offrant ainsi un environnement partagé de tenant, ou multitenant. Pour faciliter le processus, Access Zones associe les trois principaux composants d'accès externes :

- Configuration du réseau du cluster
- Protocole d'accès aux fichiers
- Authentification

Les zones SmartConnect sont associées à un ensemble de partages SMB, d'exportations NFS, de racks HDFS et à un ou plusieurs fournisseurs d'authentification pour le contrôle d'accès. Cela offre les avantages d'un système de fichiers unique géré de manière centralisée, qui peut être provisionné et sécurisé pour plusieurs tenants. Cette fonction est particulièrement utile pour les environnements d'entreprise où les multiples entités distinctes font appel à un département IT central. Autre exemple : durant une initiative de consolidation de serveurs, lors de la fusion de plusieurs serveurs de fichiers Windows associés à des forêts Active Directory séparées et non fiables.

Avec Access Zones, la zone d'accès système intégrée contient une instance de chaque fournisseur d'authentification pris en charge, tous les partages SMB disponibles et toutes les exportations NFS disponibles par défaut.

Ces fournisseurs d'authentification peuvent inclure plusieurs instances de Microsoft Active Directory, LDAP, NIS et des bases de données d'utilisateurs ou de groupes locaux.

## Administration basée sur des rôles

L'administration basée sur des rôles est un système de contrôle d'accès basé sur des rôles de gestion de cluster (RBAC) qui divise les privilèges des utilisateurs root et administrateur en privilèges plus granulaires et qui autorise l'affectation de ces rôles spécifiques. Il est possible d'accorder ces rôles à des utilisateurs sans privilèges. Par exemple, le personnel en charge des opérations du datacenter peut se voir octroyer des droits en lecture seule sur l'ensemble du cluster, ce qui lui donne un accès avec contrôle total mais ne lui permet pas de modifier la configuration. OneFS propose un ensemble de rôles prédéfinis, comprenant l'audit, l'administrateur système et sécurité, avec en plus la possibilité de créer des rôles personnalisés par zone d'accès ou dans le cluster. L'administration basée sur les rôles est intégrée avec l'interface CLI, l'interface WebUI et l'API de plate-forme de OneFS.

📖 Pour plus d'informations sur la gestion des identités, l'authentification et le contrôle d'accès dans les environnements NFS et SMB combinés, reportez-vous au guide de sécurité multiprotocole OneFS.

## Audit OneFS

OneFS offre la possibilité d'auditer la configuration système et l'activité de protocole NFS, SMB et HDFS sur un cluster. Cela permet aux entreprises de répondre à diverses obligations de gouvernance des données et de conformité légale auxquelles elles peuvent être liées.

Toutes les données d'audit sont stockées et protégées dans le système de fichiers du cluster, et organisées en rubriques d'audit. À partir de là, les données d'audit peuvent être exportées au travers du cadre Dell EMC Common Event Enabler (CEE) vers des applications tierces telles que Varonis DatAdvantage et Symantec Data Insight. L'audit de protocole OneFS peut être activé par zone d'accès, ce qui permet un contrôle granulaire sur l'ensemble du cluster.

Un cluster peut écrire des événements d'audit sur un maximum de cinq serveurs CEE par nœud dans une configuration parallèle avec équilibrage de charge. Cela permet à OneFS de fournir une solution d'audit d'entreprise de bout en bout.

📖 Pour plus d'informations, reportez-vous au livre blanc OneFS Audit.

## Mise à niveau des logiciels

La mise à niveau vers la dernière version de OneFS vous permet de tirer parti des nouvelles fonctionnalités et correctifs. Les clusters peuvent être mis à niveau à l'aide de deux méthodes : mise à niveau simultanée ou mise à niveau propagée.

### Mise à niveau simultanée

Une mise à niveau simultanée installe le nouveau système d'exploitation et redémarre tous les nœuds du cluster en même temps. Cette opération requiert une interruption de service temporaire (moins de 2 minutes) pendant le processus de mise à niveau, tandis que les nœuds sont redémarrés.

### Mise à niveau consécutive

Une mise à niveau consécutive met à niveau et redémarre individuellement chaque nœud du cluster de façon séquentielle. Dans le cadre d'une mise à niveau consécutive, le cluster reste en ligne et continue de servir les données aux clients sans aucune interruption de service. Avant OneFS 8.0, une mise à niveau ne pouvait être exécutée qu'avec une même famille de versions de code OneFS, et non avec différentes révisions de versions de code majeures OneFS. À partir de OneFS 8.0, chaque nouvelle version est mise à niveau à partir de la version précédente.

### Mises à niveau sans perturbation

Les mises à niveau sans perturbation permettent à un administrateur de cluster de mettre à niveau le système d'exploitation de stockage, tandis que les utilisateurs finaux continuent à accéder aux données sans aucune erreur ni interruption. La mise à jour du système d'exploitation sur un cluster est une question simple de mise à niveau consécutive. Au cours de ce processus, un seul nœud est mis à niveau vers le nouveau code, et les clients NFS et SMB3 actifs qui y sont rattachés sont automatiquement migrés vers d'autres nœuds du cluster. La mise à niveau partielle est également autorisée, ce qui permet de mettre à niveau un sous-ensemble de nœuds de cluster. Le sous-ensemble de nœuds peut aussi être développé lors de la mise à niveau. Une mise à niveau peut être interrompue et reprise, ce qui permet aux clients de répartir les mises à niveau sur plusieurs fenêtres de maintenance plus petites. De plus, OneFS 8.2.2 et les versions ultérieures offrent des mises à niveau parallèles, par lesquelles les clusters peuvent mettre à niveau un voisinage entier, ou un domaine de pannes, à la fois, ce qui réduit considérablement la durée des mises à niveau de clusters volumineux. OneFS 9.2 et les versions ultérieures associent les mises à niveau du système d'exploitation et du firmware, ce qui réduit considérablement l'impact et la durée des mises à niveau en leur permettant de se produire simultanément. Les versions 9.2 et ultérieures incluent également des mises à niveau basées sur la maintenance, qui empêchent les nœuds de redémarrer ou de relancer les services de protocole tant que tous les clients SMB ne sont pas déconnectés du nœud.

**Retour arrière possible**

OneFS prend en charge le retour arrière de la mise à niveau, ce qui permet de renvoyer un cluster avec une mise à niveau non validée vers sa version précédente de OneFS.

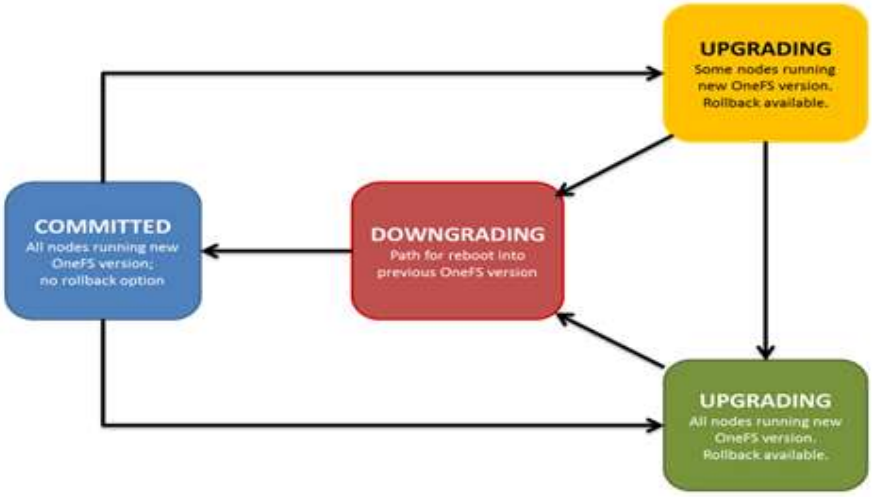


Figure 28 : États de mise à niveau sans perturbation OneFS

**Mises à jour automatiques du micrologiciel**

Les clusters OneFS prennent en charge les mises à jour automatiques des firmwares de disque pour les nouveaux disques et les disques de remplacement, dans le cadre du processus de mise à jour de firmwares sans perturbation. Les mises à jour du micrologiciel sont fournies par le biais de packages de prise en charge des disques, qui simplifient et rationalisent la gestion des disques existants et nouveaux sur l'ensemble du cluster. Cela permet de s'assurer que le micrologiciel du disque est à jour et limite la probabilité des défaillances en cas de problème de disque connu. Par conséquent, les mises à jour automatiques des firmwares de disque constituent un composant important de la stratégie de haute disponibilité et des opérations sans perturbation d'OneFS. Le micrologiciel de disque et de nœud peut être appliqué sous la forme d'une mise à niveau consécutive ou d'un redémarrage complet du cluster.

Avant OneFS 8.2, les mises à jour de firmwares de nœud devaient être installées sur un nœud à la fois. L'opération était donc fastidieuse, en particulier dans les grands clusters. Les mises à jour du micrologiciel des nœuds peuvent désormais être organisées sur un cluster par le biais d'une liste de nœuds à mettre à jour simultanément. L'outil d'aide à la mise à niveau peut être utilisé pour sélectionner une combinaison de nœuds à mettre à jour simultanément et une liste explicite de nœuds à ne pas mettre à jour simultanément (par exemple, les nœuds d'une paire de nœuds).

**Réalisation de la mise à niveau**

Dans le cadre d'une mise à niveau, OneFS exécute automatiquement un contrôle de vérification préalable à l'installation. Cela permet de vérifier que la configuration de l'installation actuelle de OneFS est compatible avec la version de OneFS qui doit être mise à niveau. Lorsqu'une configuration non prise en charge est trouvée, la mise à niveau est arrêtée et des instructions de dépannage s'affichent. L'exécution proactive de la vérification de préinstallation avant de lancer une mise à niveau permet d'éviter toute interruption due à une configuration incompatible.

# Logiciel de gestion et de protection des données OneFS

OneFS propose un portefeuille complet de logiciels de gestion et de protection des données pour répondre à vos besoins :

| Module logiciel | Fonction | Description |
|---|---|---|
| CloudIQ™ | **Surveiller l'intégrité du cluster** | Implémentation d'une analytique prédictive et intelligente pour surveiller proactivement l'intégrité de votre cluster. |
| InsightIQ™ | **Gestion des performances** | Optimisation des performances de votre cluster au moyen d'outils innovants de monitoring et reporting. |
| DataIQ™ | **Gestion et analyse des données** | Recherchez des données, accédez-y et gérez-les en quelques secondes, où qu'elles se trouvent : sur le stockage en mode fichier et objet, sur site ou dans le Cloud. Bénéficiez d'une vue d'ensemble sur les systèmes de stockage hétérogènes à l'aide d'une seule et même interface, en supprimant efficacement les données piégées dans des silos. |
| SmartPools™ | **Gestion des ressources** | Mise en œuvre d'une hiérarchisation automatisée du stockage extrêmement efficace pour optimiser les performances et les coûts de stockage. |
| SmartQuotas™ | **Gestion des données** | Attribution et gestion de quotas, permettant de partitionner et d'allouer de manière dynamique et transparente l'espace de stockage sous forme de segments faciles à gérer aux niveaux cluster, répertoire, sous-répertoire, utilisateur et groupe |
| SmartConnect™ | **Accès aux données** | Activation de l'équilibrage de la charge des connexions client et basculement NFS sur incident et retour arrière dynamiques de ces connexions sur l'ensemble des nœuds de stockage pour optimiser l'utilisation des ressources du cluster. |
| SnapshotIQ™ | **Protection des données** | Protection des données de manière fiable et efficace à l'aide de snapshots quasi instantanés avec un impact faible ou nul sur les performances. Restauration rapide des données critiques grâce à des restaurations de snapshot à la demande et quasi immédiates. Créez des copies modifiables et économes en taille et en durée d'un snapshot en lecture seule avec des snapshots en création OneFS. |
| SyncIQ™ | **Réplication des données** | Réplication et distribution asynchrones de volumineux Datasets critiques sur plusieurs systèmes de stockage partagés, répartis sur différents sites, pour une fonction de reprise après sinistre fiable. Simplicité des fonctions de basculement sur incident et de retour arrière pour améliorer la disponibilité des données critiques. |
| SmartLock™ | **Rétention de données** | Protection des données critiques en cas de suppression ou de modification accidentelle, prématurée ou malveillante grâce à notre solution logicielle de type WORM (Write Once Read Many), et respect des exigences de conformité et de gouvernance les plus strictes (norme SEC 17a-4, par exemple). |
| SmartDedupe™ | **Déduplication des données** | Optimisation de l'efficacité du stockage par l'analyse du cluster pour rechercher les blocs identiques, puis par la suppression des réplicas, ce qui réduit la quantité de stockage physique requise. |
| CloudPools™ | **Hiérarchisation sur le Cloud** | CloudPools vous permet de définir les données de votre cluster qui doivent être archivées sur le stockage Cloud. Les fournisseurs Cloud incluent Microsoft Azure, Google Cloud, Amazon S3, Dell EMC ECS et OneFS natif. |

Table 3 : Gamme de services de données Dell EMC PowerScale

Reportez-vous à la documentation du produit pour plus d'informations.

# Conclusion

Avec les solutions NAS scale-out Dell EMC optimisées par le système d'exploitation OneFS, les organisations peuvent évoluer de téraoctets en pétaoctets dans un seul système de fichiers et un seul volume, avec un point d'administration unique. OneFS offre de hautes performances et/ou un débit élevé, sans augmenter la complexité de la gestion.

Les datacenters de nouvelle génération doivent être conçus pour une évolutivité durable. Ils doivent maîtriser la puissance de l'automatisation, tirer parti de l'uniformisation du matériel, garantir la consommation totale du fabric réseau et offrir une flexibilité optimale aux entreprises soucieuses de satisfaire un ensemble d'exigences en constante évolution.

OneFS est le système de fichiers de nouvelle génération conçu pour relever ces défis. OneFS offre les avantages suivants :

- système de fichiers unique entièrement distribué ;
- performances optimales, cluster 100 % symétrique ;
- répartition des fichiers sur tous les nœuds du cluster ;
- logiciel automatisé pour éliminer la complexité ;
- équilibrage dynamique du contenu ;
- protection flexible des données ;
- Haute disponibilité
- administration via interface Web et CLI.

OneFS est particulièrement adapté aux applications de « Big Data » basées sur des fichiers ou des données non structurées dans des environnements Data Lake d'entreprise (répertoires personnels à grande échelle, partages de fichiers, archives, virtualisation, analytique métier), mais aussi dans divers environnements informatiques hautes performances utilisant les données de façon intensive (exploration des ressources énergétiques, services financiers, services Internet et d'hébergement, business intelligence, ingénierie, fabrication, multimédia et divertissement, bioinformatique, recherche scientifique, etc.).

## ÉTAPE SUIVANTE

Contactez un agent commercial  ou un revendeur agréé Dell EMC pour découvrir les avantages des solutions de stockage NAS PowerScale pour votre entreprise.

Visitez Dell EMC PowerScale pour comparer les fonctionnalités et obtenir plus d'informations.

En savoir plus sur les solutions Dell EMC PowerScale

Contacter un expert Dell EMC

Afficher plus de ressources

Prenez part à la discussion avec #DellEMCStorage