DELL Technologies

Livre blanc technique

# Dell EMC ECS : Conception de la haute disponibilité


## Résumé

Ce document décrit les détails architecturaux de la plate-forme Dell EMC™ ECS et les moyens avec lesquels il assure la disponibilité de l'entreprise.



Juin 2021

H16344.6

# Révisions

| Date | Description |
|---|---|
| Juillet 2017 | version initiale |
| Août 2017 | Mise à jour pour inclure le contenu de la version 3.1 d'ECS |
| Mars 2019 | Mise à jour pour inclure le contenu de la version 3.3 d'ECS |
| Avril 2020 | Mise à jour des « Démarches à suivre en cas d'interruption temporaire du site (TSO) avec accès pendant l'interruption » |
| Décembre 2020 | Mise à jour de la méthode de protection des métadonnées. |
| Juin 2021 | Mise à jour pour inclure le contenu de la version 3.6.1 d'ECS |

# Remerciements

Ce livre blanc a été conçu par les éléments suivants :

Auteur : Zhu, Jarvis

# Tableau des matières

# Synthèse

Les entreprises stockent des quantités croissantes de données qu'il est essentiel de garder disponibles. Les coûts et les complexités liés à la restauration d'énormes quantités de données en cas de défaillance du système ou du site peuvent être considérables pour un département informatique.

La plate-forme Dell EMC™ ECS™ a été conçue pour répondre aux besoins de capacité et de disponibilité des entreprises d'aujourd'hui. ECS offre un exaoctet d'évolutivité avec la prise en charge d'une infrastructure d'objets distribuée à l'échelle mondiale. Il est conçu pour assurer la disponibilité de l'entreprise avec la détection automatique des pannes et les options d'auto-restauration intégrées.

Ce document décrit les détails architecturaux d'ECS et les moyens avec lesquels il assure la disponibilité de l'entreprise. Il aborde les particularités telles que :

- La manière dont l'infrastructure distribuée améliore la disponibilité du système
- Les méthodes de protection des données avancées qui assurent la durabilité des données
- La distribution des données pour une disponibilité optimale
- La détection automatique des problèmes
- Les méthodes d'autoréparation intégrées
- Les détails de la résolution des pannes de disque, de nœud et de réseau
- La reprise après sinistre :
  - Les méthodes de protection d'ECS contre les pannes à l'échelle du site
  - Le mode de maintien de la cohérence dans une configuration multisite actif-actif
  - La méthode de détection des défaillances à l'échelle du site
  - Les options d'accès lors d'une panne de site
  - La manière dont la durabilité des données est rétablie après une défaillance permanente à l'échelle du site

# Terminologie

**Datacenter virtuel (VDC) :** Dans ce livre blanc, le terme datacenter virtuel (VDC) est synonyme de site ou de zone. Les ressources ECS d'un seul VDC doivent faire partie du même réseau de gestion interne.

**Géofédération :** Vous pouvez déployer le logiciel ECS dans plusieurs datacenters afin de créer une géofédération. Dans une géofédération, ECS se comporte comme une fédération librement couplée de VDC autonomes. La fédération de sites implique la fourniture de points de terminaison de réplication et de gestion pour la communication entre les sites. Une fois les sites fédérés, ils peuvent être gérés comme une infrastructure unique à partir de n'importe quel nœud de la fédération.

**Groupes de réplication :** Les groupes de réplication définissent l'endroit où les données sont protégées. Un groupe de réplications locales contient un seul VDC et protège les données d'un même VDC contre les pannes de disque ou de nœud. Les groupes de réplication globaux contiennent plusieurs VDC et protègent les données contre les pannes de disque, de nœud et de site. Les groupes de réplication sont attribués au niveau du bucket.

# 1 Présentation de la notion de haute disponibilité

La haute disponibilité peut être décrite dans deux domaines principaux : la disponibilité du système et la durabilité des données. Un système est disponible lorsqu'il peut répondre à une demande du client. La durabilité des données est assurée indépendamment de la disponibilité du système et garantit le stockage des données dans le système sans perte ni corruption. Cela signifie que même si le système ECS hors service (par exemple, en cas de panne réseau), les données sont toujours protégées.

L'architecture distribuée d'ECS assure la disponibilité du système en permettant à n'importe quel nœud d'un datacenter virtuel (VDC) ou site de répondre aux demandes des clients. Si un nœud tombe en panne, le client peut être redirigé manuellement ou automatiquement (par exemple, à l'aide de DNS ou d'un équilibreur de charge) vers un autre nœud capable de répondre à la demande.

ECS utilise une combinaison de triple mise en miroir et de codage d'effacement pour écrire des données de manière distribuée afin d'être résilient contre les pannes de disques et de nœuds. ECS prend en charge la réplication entre les sites pour augmenter la disponibilité et la résilience en protégeant contre les pannes à l'échelle du site. ECS inclut également des contrôles systématiques réguliers de l'intégrité des données avec une fonctionnalité d'autoréparation.

Lorsque l'on s'intéresse à la haute disponibilité, il est d'abord important de comprendre l'architecture et la façon dont les données sont distribuées pour une disponibilité et des performances optimales au sein d'ECS.

## 1.1 Fragments

Un fragment est un conteneur logique qu'ECS utilise pour stocker tous les types de données, y compris les données d'objet, les métadonnées fournies par le client personnalisé et les métadonnées du système ECS. Les fragments contiennent 128 Mo de données composées d'un ou plusieurs objets d'un seul bucket, comme illustré dans la Figure 1.

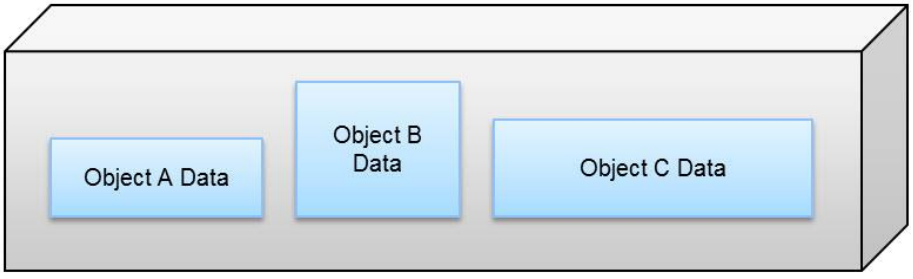


Figure 1 Fragment logique

ECS utilise l'indexation pour effectuer le suivi de toutes les données au sein d'un fragment. Pour plus d'informations, reportez-vous à la section1.2.

## 1.2 Métadonnées ECS

ECS conserve ses propres métadonnées qui assurent le suivi de l'emplacement des données, ainsi que de l'historique des transactions. Ces métadonnées sont conservées dans les tables logiques et les journaux.

Les tables contiennent des paires clé-valeur pour stocker les informations relatives aux objets. Une fonction de hachage est utilisée pour effectuer des recherches rapides de valeurs associées à une clé. Ces paires clé-valeur sont stockées dans l'arborescence B+ pour une indexation rapide des emplacements de données. En stockant la paire clé-valeur dans un arbre de recherche équilibré comme l'arborescence B+, l'emplacement des données et des métadonnées est accessible rapidement. De plus, pour améliorer davantage les performances de requête de ces tables logiques, ECS implémente une arborescence de fusion à deux niveaux, structurée à la manière d'un journal (LSM). Ainsi, deux structures d'arbre cohabitent : une arborescence de petite taille se trouve dans la mémoire (table de mémoire) et l'arborescence B+ principale réside sur le disque. Par conséquent, la recherche de paires clé-valeur interroge d'abord la table de mémoire et, si la valeur n'est pas en mémoire, elle interrogera l'arborescence B+ principale sur le disque.

L'historique des transactions est enregistré dans les journaux log et ces logs sont écrits sur les disques. Les journaux suivent les transactions qui n'ont pas encore été validées et inscrites dans l'arborescence B+. Après la consignation d'une transaction dans un journal, la table en mémoire est mise à jour. Une fois que la table en mémoire est saturée ou après une période définie, la table est fusionnée, triée ou vidée dans l'arborescence B+ sur le disque, et un point de contrôle est enregistré dans le journal. Ce processus est illustré dans la Figure 2.

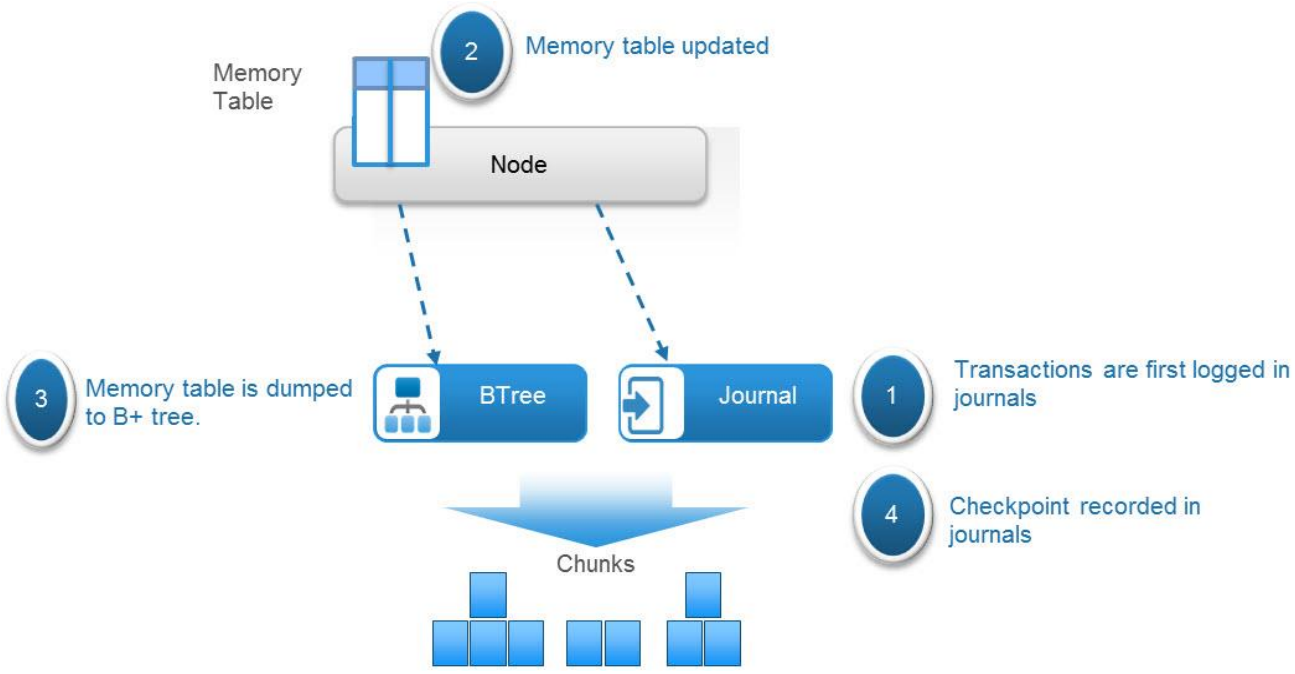


Figure 2 Workflow des mises à jour des transactions dans les tables ECS

Les journaux et les arborescences B+ sont écrits en fragments.

ECS utilise plusieurs tables différentes, chacune d'elles pouvant devenir très volumineuses. Afin d'optimiser les performances des recherches de tables, chaque table est divisée en partitions réparties sur les nœuds d'un VDC ou d'un site. Le nœud dans lequel la partition est écrite devient alors le propriétaire/l'autorité de cette partition ou section de la table.

L'un d'entre eux est une table de fragments qui assure le suivi de l'emplacement physique des fragments et des copies de réplica sur le disque. La Tableau 1 présente un exemple de partition de la table de fragments qui, pour chaque fragment, identifie son emplacement physique en répertoriant le disque dans le nœud, le fichier dans le disque, le décalage au sein de ce fichier et la longueur des données. Ici, nous pouvons voir que l'ID de fragment C1 est un code d'effacement et que l'ID de fragment C2 est mis en miroir triple. Vous trouverez plus d'informations sur la mise en miroir triple et le codage d'effacement dans la section 1.4 de ce document.

Tableau 1 Exemple de partition de table de fragments

| ID de fragment | Emplacement du fragment |
|---|---|
| C1 | Node1:Disk1:file1:offset1:length<br>Node2:Disk1:File1:offset1:length<br>Node3:Disk1:File1:offset1:length<br>Node4:Disk1:File1:offset1:length<br>Node5:Disk1:File1:offset1:length<br>Node6:Disk1:File1:offset1:length<br>Node7:Disk1:File1:offset1:length<br>Node8:Disk1:File1:offset1:length<br>Node1:Disk2:File1:offset1:length<br>Node2:Disk2:File1:offset1:length<br>Node3:Disk2:File1:offset1:length<br>Node4:Disk2:File1:offset1:length<br>Node5:Disk2:File1:offset1:length<br>Node6:Disk2:File1:offset1:length<br>Node7:Disk2:File1:offset1:length<br>Node8:Disk2:File1:offset1:length |
| C2 | Node1:Disk3:File1:offset1:length<br>Node2:Disk3:File1:offset1:length<br>Node3:Disk3:File1:offset1:length |

La table d'objets est un autre exemple, elle est utilisée pour le mappage de fragment des noms d'objet. La Tableau 2 présente un exemple de partition d'une table d'objets qui détaille le ou les fragments et l'emplacement d'un objet dans le fragment.

Tableau 2 Exemple de table d'objets

| Nom de l'objet | ID de fragment |
|---|---|
| ImgA | C1:offset:length |
| FileA | C4:offset:length<br>C6:offset:length |

Le mappage des propriétaires de partitions des tables est géré par un service, appelé vnest, qui s'exécute sur tous les nœuds. La Tableau 3 montre un exemple d'une partie d'une table de mappage vnest.

Tableau 3 Exemple de table de mappage vnest

| ID de la table | Propriétaire de la partition de table |
|---|---|
| Table 0 P1 | Nœud 1 |

| ID de la table | Propriétaire de la partition de table |
|---|---|
| Table 0 P2 | Nœud 2 |

## 1.3 Domaine de panne

En règle générale, les domaines de panne sont liés à une conception d'ingénierie qui prend en compte les composants d'une solution qui présentent un risque de défaillance. Le logiciel ECS reconnaît automatiquement quels disques se trouvent dans le même nœud et quels nœuds se trouvent dans le même rack. Afin de se protéger contre la plupart des scénarios de panne, le logiciel ECS est conçu pour utiliser ces informations lors de l'écriture de données. Les principes de base des domaines de panne qu'ECS utilise sont les suivants :

- ECS n'écrit jamais les éléments d'un même fragment sur le même disque d'un nœud
- ECS répartit les éléments d'un fragment de manière égale sur les nœuds
- ECS reconnaît rack, si un VDC ou un site contient plusieurs racks, en supposant qu'il dispose d'un espace suffisant, ECS déploie les meilleurs efforts pour répartir équitablement les éléments d'un fragment sur ces racks.

## 1.4 Méthodes avancées en matière de protection des données

Dans ECS, lorsqu'un objet est créé, il comprend des données d'écriture, des métadonnées personnalisées et des métadonnées ECS. Les métadonnées ECS incluent des fragments de journal et des fragments btree. Chacun d'eux est écrit dans un fragment logique différent qui contiendra environ 128 Mo de données d'un ou plusieurs objets. ECS utilise une combinaison de triple mise en miroir et de codage d'effacement pour protéger les données au sein d'un datacenter virtuel (VDC) ou d'un site.

- La mise en miroir triple garantit l'écriture de trois copies de données, ce qui protège les données contre les pannes de deux nœuds.
- Le codage d'effacement offre une meilleure protection des données contre les pannes de disques et de nœuds. Il utilise le schéma de codage d'effacement Reed Solomon qui divise les fragments en données et fragments de codage qui sont répartis équitablement entre les nœuds au sein d'un VDC ou d'un site.

En fonction de la taille et du type de données, elles seront écrites à l'aide de l'une des méthodes de protection des données indiquées dans la Tableau 4.

Tableau 4 Déterminer quel niveau de protection des données sera utilisé pour différents types de données

| Type de données | Méthode de protection des données utilisée |
|---|---|
| Fragments de journal | Mise en miroir triple |
| Fragments d'arborescence B / métadonnées personnalisées | Codage d'effacement avec segments de données redondantes |
| Données d'objet <128 Mo | Mise en miroir triple et codage d'effacement en place |
| Données d'objet >128 Mo | Codage d'effacement dans la ligne |

**Remarque :** Dans l'architecture All-Flash comme EXF900, la protection des fragments d'arborescence B consiste en une mise en miroir triple

### 1.4.1 Mise en miroir triple

La méthode d'écriture en triple miroir s'applique aux fragments de journal ECS, dont ECS crée trois copies de réplica. Chaque copie de réplica est écrite sur un seul disque sur différents nœuds dans les domaines de panne. Cette méthode protège les données de fragments contre les pannes touchant deux nœuds ou deux disques.

La Figure 3 montre un exemple de mise en miroir triple par laquelle un fragment logique, contenant 128 Mo de métadonnées, dispose de trois copies de réplica, chacune écrite sur un nœud différent.

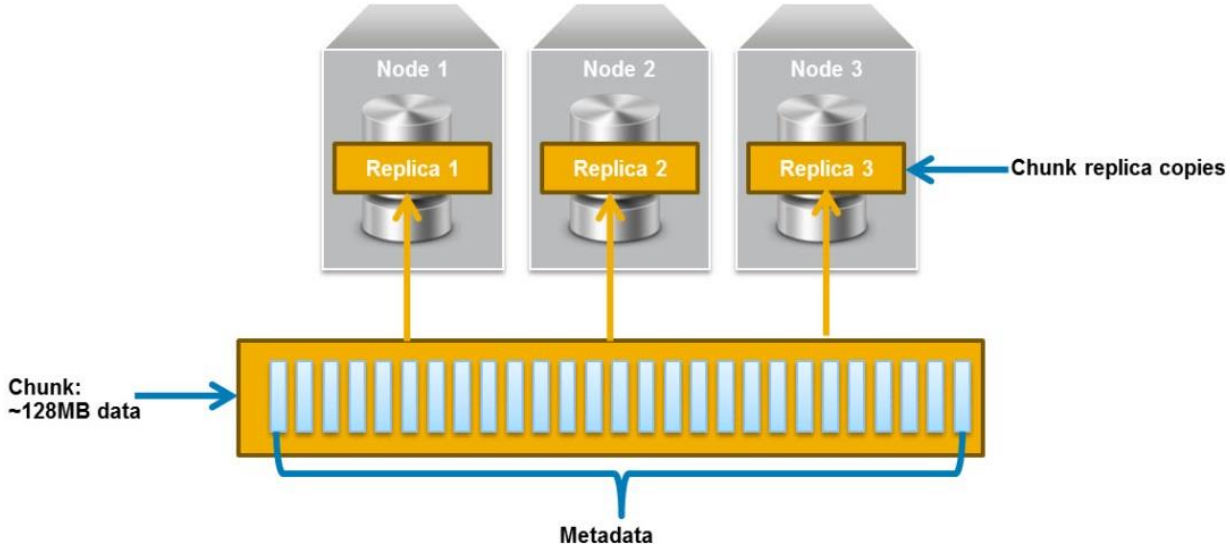


Figure 3 Mise en miroir triple

### 1.4.2 Codage d'effacement avec segments de données redondantes

Le codage d'effacement avec la méthode d'écriture des segments de données redondantes s'applique aux fragments d'arborescence B ECS et aux métadonnées d'objet personnalisées. Il comprend 12 segments de données, 12 segments de données répliqués et 4 segments de parité. Le nouveau schéma en arborescence B des fragments de données redondantes EC permet d'économiser la charge de protection des métadonnées.

### 1.4.3 Mise en miroir triple et codage d'effacement en place

Cette méthode d'écriture s'applique aux données de tout objet dont la taille est inférieure à 128 Mo.

Lorsqu'un objet est créé, il est écrit dans un fragment, ce qui permet à ECS de créer trois copies de réplica comme suit :

- Une copie est écrite dans des fragments répartis sur différents nœuds et disques. La distribution répartit les fragments sur autant de domaines de panne que possible. La taille varie sur chaque disque, et dépend du schéma de codage d'effacement utilisé.
  - Si le schéma de codage d'effacement est la valeur par défaut (12+4), chaque disque obtient un maximum d'environ 10,67 Mo
  - Si le schéma de codage d'effacement est un stockage à froid (10+2), chaque disque obtient un maximum d'environ 12,8 Mo
- Une deuxième copie de réplica du fragment est écrite sur un seul disque d'un nœud.
- Une troisième copie de réplica du fragment est écrite sur un seul disque d'un autre nœud.

Cette méthode offre une mise en miroir triple protège les données de fragments contre les pannes touchant deux nœuds ou deux disques.

Des objets supplémentaires seront écrits dans le même fragment jusqu'à ce qu'il contienne environ 128 Mo de données ou après un délai prédéfini, le délai le plus court étant retenu. À ce stade, le schéma de codage d'effacement Reed Solomon calcule les fragments de codage (de parité) du fragment et les écrit sur différents disques. Cela garantit que tous les éléments d'un fragment, y compris les fragments de codage, seront écrits sur différents disques et distribués entre les domaines de panne.

Une fois que les fragments de codage ont été écrits sur le disque, les deuxième et troisième copies de réplica sont supprimées du disque. Une fois cette opération terminée, le fragment est protégé par le codage d'effacement, qui offre des niveaux de disponibilité plus élevés que ceux de la mise en miroir triple.

### 1.4.4 Codage d'effacement dans la ligne

Cette méthode d'écriture s'applique aux données de tout objet dont la taille est d'au moins 128 Mo. Les objets sont divisés en fragments de 128 Mo. Le schéma de codage d'effacement Reed Solomon calcule les fragments de codage (de parité) de chaque fragment. Chaque fragment sera écrit sur un disque différent, et distribué entre les domaines de panne. La taille varie sur chaque disque, et dépend du schéma de codage d'effacement utilisé.

- Si le schéma de codage d'effacement est la valeur par défaut (12+4), les fragments seront répartis sur 16 disques, et chaque fragment fera environ 10,67 Mo
- Si le schéma de codage d'effacement est un stockage à froid (10+ 2), les fragments seront répartis sur 12 disques, et chaque fragment fera environ 12,8 Mo

Toute partie restante d'un objet de moins de 128 Mo sera écrite à l'aide de la mise en miroir triple et du schéma de codage d'effacement en place mentionné précédemment. Par exemple, si un objet pèse 150 Mo, 128 Mo seront écrits à l'aide du codage d'effacement dans la ligne, les 22 Mo restants seront écrits à l'aide d'un codage d'effacement en mise en miroir triple et celui en place.

La Figure 4 montre un exemple de la manière dont les fragments sont répartis entre les domaines de panne. Cet exemple comporte un seul VDC ou site qui s'étend sur deux racks, chaque rack contenant quatre nœuds.

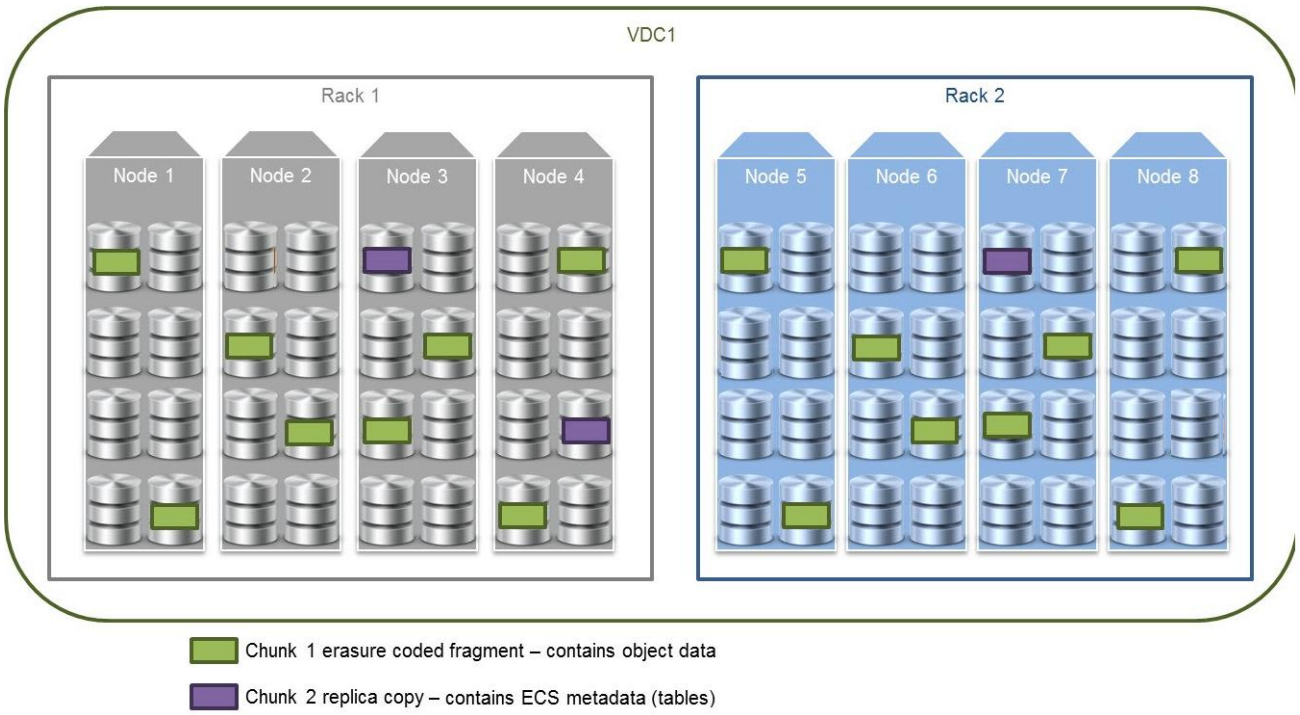


Figure 4 Répartition des fragments entre les domaines de panne

- Le fragment 1 contient des données d'objet qui ont été codées à l'aide d'un codage d'effacement de 12+4. Les éléments sont répartis uniformément sur les 8 nœuds, quatre par rack. Chaque nœud contient deux fragments qu'il écrit sur deux disques différents.
- Le fragment 2 contient des métadonnées ECS (tables) et est donc mis en miroir triple. Chaque copie de réplica est écrite sur un nœud différent, chacune sur un seul disque. Les copies s'étendent sur des racks pour offrir la disponibilité la plus élevée.

## 1.5 Niveaux de protection du codage d'effacement

En fonction du schéma de codage d'effacement sélectionné lors de la création du pool de stockage, les données codées par effacement sont protégées contre les pannes suivantes.

### 1.5.1 Codage d'effacement par défaut (12+4) :

ECS nécessite un minimum de quatre nœuds pour pouvoir effectuer le codage d'effacement en utilisant l'option par défaut. Le codage d'effacement s'arrête si un pool de stockage contient moins de quatre nœuds, ce qui signifie que le niveau de protection sera assuré par la mise en miroir triple. Pendant ce temps, les trois copies de réplica seront conservées et la parité n'est calculée sur aucun fragment. Une fois que des nœuds supplémentaires sont ajoutés au pool de stockage et respectent le nombre minimal de nœuds pris en charge, le codage d'effacement se poursuit sur ces fragments ainsi que sur les nouveaux fragments.

Pour chaque fragment de 128 Mo, le codage d'effacement par défaut écrit douze fragments de données et quatre fragments de codage, chacun d'une taille d'environ 10,67 Mo. Il protège les données de fragments contre la perte d'un maximum de quatre éléments d'un fragment, ce qui peut inclure les scénarios de défaillance suivants illustrés dans la Tableau 5.

Tableau 5 Protection par codage d'effacement par défaut

| Nombre de nœuds dans le VDC | Nombre d'éléments de fragments par nœud | Les données codées par effacement sont protégées contre |
|---|---|---|
| 5 nœuds | 4 | • Perte de quatre disques max. ou<br>• Perte d'un nœud |
| 6 ou 7 nœuds | 3 | • Perte de quatre disques max. ou<br>• Perte d'un nœud et d'un disque d'un second nœud |
| 8 nœuds ou plus | 2 | • Perte de quatre disques max. ou<br>• Perte de deux nœuds ou<br>• Perte d'un nœud et de deux disques |
| 16 nœuds ou plus | 1 | • Perte de quatre nœuds ou<br>• Perte de trois nœuds et des disques d'un nœud supplémentaire ou<br>• Perte de deux nœuds et des disques de deux nœuds différents ou<br>• Perte d'un nœud et des disques d'au maximum trois nœuds différents ou<br>• Perte de quatre disques de quatre nœuds différents |

**Remarque :** La Tableau 5 indique les niveaux de protection possibles avec une distribution complète des éléments de fragments. Il peut y avoir des cas où il existe plus de fragments sur un nœud, par exemple si l'espace disponible d'un nœud est insuffisant. Dans ce cas, les niveaux de protection peuvent varier.

### 1.5.2 Codage d'effacement par stockage à froid (10+2) :

ECS nécessite un minimum de six nœuds pour pouvoir effectuer le codage d'effacement en utilisant l'option par stockage à froid. Le codage d'effacement s'arrête si un pool de stockage contient moins de six nœuds, ce qui signifie que les trois copies de réplica seront conservées et que la parité ne sera pas calculée sur un fragment. Une fois que des nœuds supplémentaires sont ajoutés au pool de stockage, le codage d'effacement se poursuit sur ces fragments ainsi que sur les nouveaux fragments.

Pour chaque fragment de 128 Mo, le codage d'effacement par stockage à froid écrit dix fragments de données et deux fragments de codage, chacun d'une taille d'environ 12,8 Mo. Il protège les données de fragments contre la perte d'un maximum de deux éléments d'un fragment, ce qui peut inclure les scénarios de défaillance suivants illustrés dans la Tableau 6.

Tableau 6 Protection du codage d'effacement par stockage à froid

| Nombre de nœuds dans le VDC | Nombre d'éléments de fragments par nœud | Les données codées par effacement sont protégées contre |
|---|---|---|
| 11 nœuds ou moins | 2 | • Perte de deux disques max. ou<br>• Perte d'un nœud |
| 12 nœuds ou plus | 1 | • Perte de n'importe quel nombre de disques de deux nœuds différents ou<br>• Perte de deux nœuds |

**Remarque :** La table indique les niveaux de protection possibles avec une distribution complète des éléments de fragments. Il peut y avoir des cas où il existe plus de fragments sur un nœud, par exemple si l'espace disponible d'un nœud est insuffisant. Dans ce cas, les niveaux de protection peuvent varier.

## 1.6 Sommes de contrôle

Pour garantir l'intégrité des données, ECS stocke la somme de contrôle des données écrites. Les sommes de contrôle sont effectuées par unité d'écriture, jusqu'à 2 Mo. Par conséquent, les sommes de contrôle peuvent être effectuées pour un fragment d'objet dans le cas d'écritures d'objets volumineux, ou pour chaque objet dans le cas d'écritures d'objets de petite taille de moins de 2 Mo. Lors des opérations d'écriture, la somme de contrôle est calculée en mémoire, puis écrite sur le disque. Lors des lectures, les données sont lues avec la somme de contrôle, puis la somme de contrôle est calculée en mémoire à partir de la lecture des données et comparée à la somme de contrôle stockée sur le disque pour déterminer l'intégrité des données. De plus, le moteur de stockage exécute régulièrement un vérificateur de cohérence en arrière-plan et effectue la vérification de la somme de contrôle sur l'ensemble du jeu de données.

## 1.7 Écriture d'objets

Lorsqu'une opération d'écriture se produit dans ECS, elle commence par l'envoi d'une demande à un nœud par un client. ECS a été conçu comme une architecture distribuée qui permet à n'importe quel nœud d'un VDC ou d'un site de répondre à une demande de lecture ou d'écriture. Une demande d'écriture implique l'écriture des données d'objet, des métadonnées d'objet personnalisées et l'enregistrement de la transaction dans un journal log. Une fois cette opération terminée, le client reçoit un accusé de réception.

La Figure 5 et les étapes décrites ci-dessus donnent un aperçu de haut niveau d'un workflow d'écriture.

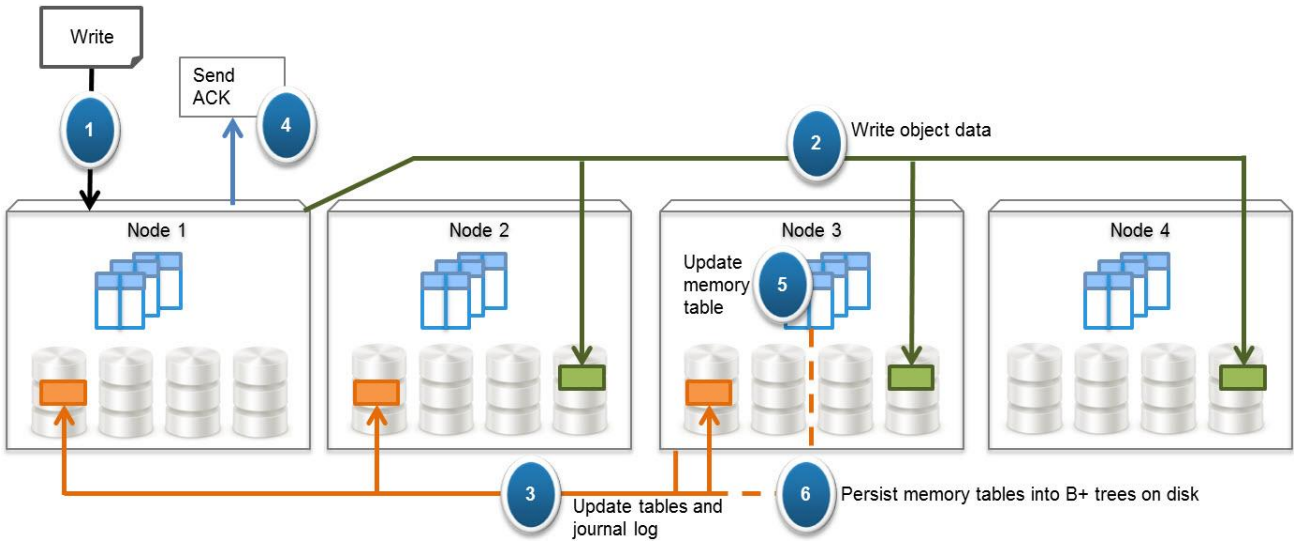


Figure 5 Workflow d'écriture d'objets

1. Une demande d'écriture d'objet est reçue. N'importe quel nœud peut répondre à cette demande, mais dans cet exemple, le Nœud 1 traite la demande.
2. En fonction de la taille de l'objet, les données sont écrites sur un ou plusieurs fragments. Chaque fragment est protégé à l'aide de schémas de protection des données avancés tels que la mise en miroir triple et le codage d'effacement. Avant d'écrire les données sur le disque, ECS exécute la fonctionnalité de somme de contrôle et conserve le résultat.

   Les données sont ajoutées à un fragment. Puisque cet objet n'a qu'une taille de 10 Mo, il utilise la mise en miroir triple et le codage d'effacement en place. Il en résulte des écritures sur trois disques de trois nœuds différents, dans cet exemple, Nœud 2, Nœud 3 et Nœud 4. Ces trois nœuds envoient des accusés de réception au Nœud 1.
3. Une fois les données de l'objet écrites avec succès, les métadonnées de l'objet sont stockées. Dans cet exemple, le Nœud 3 est propriétaire de la partition de la table d'objets dans lequel cet objet appartient. En tant que nœud propriétaire, le Nœud 3 écrit le nom de l'objet et l'ID de fragment dans cette partition des journaux log de la table d'objets. Les journaux log sont mis en miroir triple, de sorte que le Nœud 3 envoie des copies de réplica à trois nœuds différents en parallèle, dans cet exemple, Nœud 1, Nœud 2 et Nœud 3.
4. L'accusé de réception est envoyé au client.
5. Dans un processus en arrière-plan, la table de mémoire est mise à jour.
6. Une fois que la table en mémoire est saturée ou après une période définie, la table est fusionnée, triée ou vidée dans l'arborescence B+ sous forme de fragments, et un point de contrôle est enregistré dans le journal.

## 1.8 Lecture d'objets

ECS a été conçu comme une architecture distribuée qui permet à n'importe quel nœud d'un VDC ou d'un site de répondre à une demande de lecture ou d'écriture. Une demande de lecture implique la recherche de l'emplacement physique des données à l'aide de table de recherches à partir du propriétaire de l'enregistrement de partition, ainsi que les lectures de décalage d'octets, la validation de la somme de contrôle et le renvoi des données au client demandeur.

La Figure 6 et les étapes ci-dessus donnent un aperçu d'un workflow de lecture.

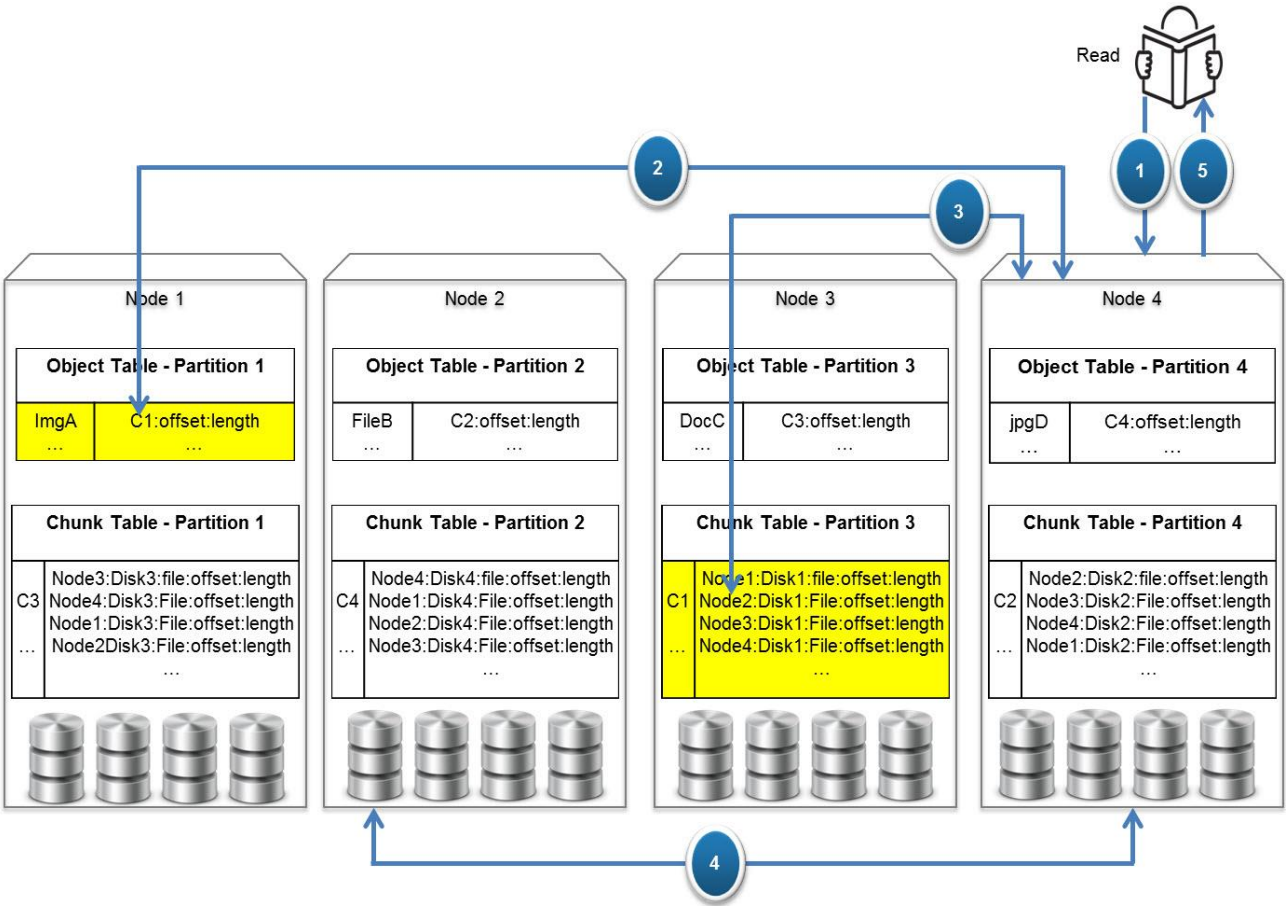


Figure 6 Workflow de lecture d'objets

1. Une demande de lecture est reçue pour ImgA. N'importe quel nœud peut répondre à cette demande, mais dans cet exemple, le Nœud 4 traite la demande.
2. Le Nœud 4 demande les informations de fragment au Nœud 1, propriétaire de la partition de table d'objets pour ImgA.
3. Sachant qu'ImgA se trouve dans C1 à un décalage et une longueur particuliers, le Nœud 4 demande l'emplacement physique du fragment au Nœud 3, propriétaire de la partition de table de fragments pour C1.
4. Maintenant que le Nœud 4 connaît l'emplacement physique d'ImgA, il demande ces données au nœud ou plusieurs nœuds qui contiennent le ou les fragments de données de ce fichier, dans cet exemple, le Nœud 2 Disque 1. Le ou les nœuds effectuent une lecture de décalage d'octet et renvoient les données au Nœud 4.
5. Le Nœud 4 valide la somme de contrôle, puis renvoie les données au client demandeur.

**Remarque :** À l'étape 4, pour l'architecture All-Flash comme EXF900, chaque nœud peut lire les données directement d'un autre nœud, contrairement aux architectures à disque dur comme EX300, EX500 et EX3000, pour lesquelles chaque nœud peut lire que les données stockées en son sein.

# 2 Disponibilité de site local

L'architecture distribuée d'ECS offre une haute disponibilité sous la forme de la disponibilité du système et de la durabilité des données en cas de pannes diverses. Cette section est consacrée à la manière dont la disponibilité est maintenue lors des pannes de site local.

## 2.1 Disk failure

La section Architecture décrit la manière dont ECS utilise une combinaison de triple mise en miroir et de codage d'effacement pour écrire des données de manière distribuée afin d'être résilient contre diverses pannes.

Pour garantir la cohérence des données, des sommes de contrôle sont validées lors des lectures et par un vérificateur de cohérence. Le vérificateur de cohérence est un processus en arrière-plan qui effectue régulièrement la vérification de la somme de contrôle sur l'ensemble du jeu de données. Les demandes de lecture exécutent également la vérification de la somme de contrôle.

Si un fragment manque à une demande de lecture, en raison d'un lecteur qui ne répond pas ou d'un échec de vérification de la somme de contrôle, une notification est envoyée au gestionnaire de fragments. Le gestionnaire de fragments lance une reconstruction du ou des fragments manquants à l'aide des données codées par effacement restantes et des fragments de parité, ou des copies de réplica, après quoi il met à jour les informations de fragment. Une fois que les fragments ont été recréés, toutes les demandes de lecture, en attente ou nouvelles, utilisent les informations de fragment mises à jour pour demander le ou les fragments de données et répondre à la demande de lecture.

Les nœuds ECS effectuent constamment des bilans de santé sur les disques qui leur sont directement rattachés. Si un disque ne répond plus, le nœud ECS notifie le gestionnaire de fragments d'arrêter de l'inclure dans les nouvelles opérations d'écriture. S'il ne répond toujours pas après un certain temps (la valeur par défaut est de 60 minutes), une notification est envoyée au gestionnaire de fragments pour recréer les données du disque en panne. Le nœud ECS identifie les fragments qui comportent des blocs sur le disque en panne et qui doit donc être restauré. Il envoie ces informations au gestionnaire de fragments qui démarre la récupération parallèle de tous les fragments stockés sur le disque en panne. Les éléments des fragments sont restaurés sur d'autres disques à l'aide des fragments codés par effacement restants ou des copies de réplica. À mesure que de nouveaux fragments sont écrits, les tables de fragments associées sont mises à jour avec les informations. Le gestionnaire de fragments supprime également les fragments du disque en panne, si c'est possible. Si le disque est de nouveau en ligne et qu'il :

- ne répond plus depuis moins d'une période définie (la valeur par défaut est de 90 minutes), les opérations de récupération restantes seront annulées.
- ne répond pas pendant une période définie (la valeur par défaut est de 90 minutes) ou s'il est signalé comme étant en échec par le gestionnaire de matériel, ECS supprime le disque. Une fois qu'un disque a été supprimé, les opérations de récupération restantes se poursuivent jusqu'à la fin. Lorsque la récupération est terminée, le gestionnaire de fragments supprime toutes les références à ce disque en panne dans le tableau des fragments.

  Si le lecteur est mis en ligne après sa suppression, il est ajouté en tant que nouveau lecteur et le gestionnaire de fragments l'inclut dans les nouvelles opérations d'écriture.

## 2.2 Défaillance d'un nœud ECS

ECS effectue constamment des bilans de santé sur les nœuds. Afin de maintenir la disponibilité du système, l'architecture distribuée ECS permet à n'importe quel nœud d'accepter les demandes client. Si un nœud est en panne, un client peut être redirigé manuellement ou automatiquement (par exemple, à l'aide de DNS ou d'un répartiteur de charge) vers un autre nœud capable de répondre à la demande.

Afin de ne pas déclencher des opérations de reconstruction pour des événements erronés, une opération de reconstruction complète n'est pas déclenchée, sauf si un nœud échoue un nombre défini de bilans de santé séquentiels, la valeur par défaut étant de 60 minutes. Si une demande d'E/S arrive pour un nœud qui ne répond pas, mais avant qu'une reconstruction complète ne soit lancée :

- Toute demande de table de partition hébergée sur un nœud qui ne répond pas déclenchera la répartition de la propriété des tables de partition demandées sur les nœuds restants du site. Une fois cette opération terminée, la demande se termine avec succès.
- Toute demande d'E/S pour des données qui existent sur les disques du nœud qui ne répond pas sera reconstruite en utilisant les données codées par effacement restantes et les fragments de parité, ou les copies de réplica, après quoi il met à jour les informations de fragment. Une fois que les fragments ont été recréés, toutes les demandes de lecture, en attente ou nouvelles, utilisent les informations de fragment mises à jour pour demander le ou les fragments de données et répondre à la demande de lecture.

Si un nœud échoue un nombre défini de bilans de santé séquentiels, la valeur par défaut étant de 60 minutes, ce nœud est considéré comme étant en panne. Cela déclenche automatiquement une opération de recréation des tables de partition et des fragments sur les disques appartenant au nœud en panne.

Dans le cadre de l'opération de recréation, une notification est envoyée au gestionnaire de fragments qui démarre une récupération parallèle de tous les fragments stockés sur les disques des nœuds en panne. Cela peut inclure des fragments contenant des données d'objet, des métadonnées personnalisées fournies par le client et des métadonnées ECS. Si le nœud en échec revient en ligne, l'état mis à jour est envoyé au gestionnaire de fragments et toutes les opérations de récupération non terminées sont annulées. Vous trouverez plus d'informations sur la récupération des éléments de fragment dans la section panne de disque ci-dessus.

Outre la surveillance matérielle, ECS surveille également tous les services et tables de données sur chaque nœud.

- Si une table tombe en panne, mais que le nœud est toujours actif, il essaiera automatiquement de réinitialiser la table sur le même nœud.
- S'il détecte une défaillance du service, il tente d'abord de redémarrer le service.

Si cela échoue, il redistribue la propriété des tables détenues par le nœud ou le service en panne sur tous les nœuds restants dans le VDC ou le site. Les modifications de propriété impliquent la mise à jour des informations vnest et la recréation des tables de mémoire détenues par le nœud en panne. Les informations sur vnest sont mises à jour sur les nœuds restants avec les informations sur le propriétaire de la nouvelle table de partition.

Les tables de mémoire du nœud en échec sont recréées en réexécutant les entrées de journal écrites après le dernier point de contrôle de journal réussi.

### 2.2.1 Défaillance de plusieurs nœuds

Il existe des scénarios où plusieurs nœuds peuvent tomber en panne au sein d'un site. Plusieurs nœuds peuvent tomber en panne soit un par un, soit simultanément.

- **Défaillance un par un :** Lorsque les nœuds tombent en panne un par un, cela signifie qu'un nœud tombe en panne, que toutes les opérations de récupération sont terminées, puis qu'un deuxième nœud tombe en panne. Cela peut se produire plusieurs fois et est similaire à un VDC qui passerait de 4 sites → 3 sites → 2 sites → 1 site. Cela nécessite que les nœuds restants disposent d'un espace suffisant pour effectuer les opérations de récupération.
- **Défaillance simultanée :** Lorsque les nœuds échouent simultanément, cela signifie que les nœuds échouent presque en même temps, ou qu'un nœud échoue avant la fin de la récupération d'un nœud précédent en échec.

L'impact de la défaillance dépend des nœuds qui tombent en panne. La Tableau 7 et la Tableau 8 décrivent le meilleur scénario de tolérance aux pannes d'un site unique en fonction du codage d'effacement et du nombre de nœuds dans un VDC.

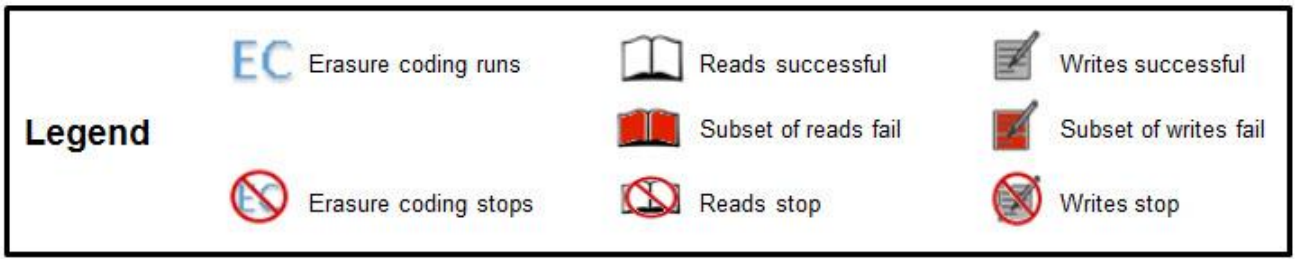


Tableau 7 Scénario idéal de défaillances de plusieurs nœuds au sein d'un site en fonction du codage d'effacement par défaut 12+4

| Nombre de nœuds dans le VDC lors de la création | Nombre total de nœuds en panne depuis la création du VDC | État après des défaillances simultanées | État après la défaillance un par un la plus récente | État actuel du VDC après les défaillances précédentes un par un |
|---|---|---|---|---|
| 5 nœuds | 1 | EC | EC | VDC à 5 nœuds avec 1 nœud défaillant |
| | 2 | | | Le VDC est passé de 5 → à 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 3 - 4 | | | Le VDC est passé de 5 → 4 → 3 ou 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| 6 nœuds | 1 | EC | EC | VDC à 6 nœuds avec 1 nœud défaillant |
| | 2 | EC | EC | Le VDC est passé de 6 → 5 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 3 | | | Le VDC est passé de 6 → 5 → 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |

| Nombre de nœuds dans le VDC lors de la création | Nombre total de nœuds en panne depuis la création du VDC | État après des défaillances simultanées | État après la défaillance un par un la plus récente | État actuel du VDC après les défaillances précédentes un par un |
|---|---|---|---|---|
| | 4 - 5 |  |  | Le VDC est passé de 6 → 5 → 4 → 3 ou 6 → 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| 8 nœuds | 1 - 2 |  |  | VDC à 8 nœuds ou le VDC est passé de 8 → 7 nœuds, 1 nœud maintenant défaillant |
| | 3 - 4 |  |  | Le VDC est passé de 8 → 7 → 6 ou 8 → 7 → 6 → 5 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 5 |  |  | Le VDC est passé de 8 → 7 → 6 → 5 → 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 6 - 7 |  |  | Le VDC est passé de 8 → 7 → 6 → 5 → 4 → 3 nœuds ou 8 → 7 → 6 → 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |

Tableau 8 Scénario idéal de défaillances de plusieurs nœuds au sein d'un site en fonction du codage d'effacement par stockage à froid 10+2

| Nombre de nœuds dans le VDC lors de la création | Nombre total de nœuds en panne depuis la création du VDC | État après des défaillances simultanées | État après la défaillance un par un la plus récente | État actuel du VDC après chaque défaillance précédente |
|---|---|---|---|---|
| 6 nœuds | 1 |  |  | VDC à 6 nœuds avec 1 nœud défaillant |
| | 2 |  |  | Le VDC est passé de 6 → 5 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 3 |  |  | Le VDC est passé de 6 → 5 → 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |

| Nombre de nœuds dans le VDC lors de la création | Nombre total de nœuds en panne depuis la création du VDC | État après des défaillances simultanées | État après la défaillance un par un la plus récente | État actuel du VDC après chaque défaillance précédente |
|---|---|---|---|---|
| | 4 - 5 |  |  | Le VDC est passé de 6 → 5 → 4 → 3 ou 6 → 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| 8 nœuds | 1 |  |  | VDC à 8 nœuds avec 1 nœud défaillant |
| | 2 |  |  | Le VDC est passé de 8 → 7 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 3 - 5 |  |  | Le VDC est passé de 8 → 7 → 6 ou 8 → 7 → 6 → 5 ou 8 → 7 → 6 → 5 → 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 6 - 7 |  |  | Le VDC est passé de 8 → 7 → 6 → 5 → 4 → 3 nœuds ou 8 → 7 → 6 → 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| 12 nœuds | 1 - 2 |  |  | VDC à 12 nœuds ou le VDC à 12 nœuds est passé de 12 → 11 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 3 - 6 |  |  | Le VDC est passé de 12 → 11 → 10 ou 12 → 11 → 10 → 9 ou 12 → 11 → 10 → 9 → 8 ou 12 → 11 → 10 → 9 → 8 → 7 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 7 - 9 |  |  | Le VDC est passé de 12 → 11 → 10 → 9 → 8 → 7 → 6 ou 12 → 11 → 10 → 9 → 8 → 7 → 6 → 5 ou 12 → 11 → 10 → 9 → 8 → 7 → 6 → 5 → 4 nœuds, 1 nœud supplémentaire est maintenant défaillant |
| | 10 - 11 |  |  | Le VDC est passé de 12 → 11 → 10 → 9 → 8 → 7 → 6 → 5 → 4 → 3 ou 12 → 11 → 10 → 9 → 8 → 7 → 6 → 5 → 4 → 3 → 2 nœuds, 1 nœud supplémentaire est maintenant défaillant |

Les règles de base pour déterminer les opérations qui échouent sur un site unique avec plusieurs défaillances de nœuds sont les suivantes :

- Si vous rencontrez au moins trois défaillances de nœud simultanées, certaines lectures et écritures échouent en raison de la perte potentielle des trois copies de réplica des fragments de métadonnées mises en miroir triple associés.
- L'écriture nécessite un minimum de trois nœuds.
- Le codage d'effacement s'arrête et les fragments codés par effacement sont convertis en triple protection en miroir si le nombre de nœuds est inférieur au minimum requis pour chaque codage d'effacement. Le codage d'effacement par défaut (12+4) nécessite 4 nœuds, si moins de 4 nœuds sont fonctionnels, le codage d'effacement s'arrête. Le codage d'effacement par stockage à froid (10+2) s'arrête si moins de 6 nœuds sont fonctionnels.
- Si le nombre de nœuds est inférieur au minimum requis pour le codage d'effacement, les fragments codés par effacement seront convertis en protection à mise en miroir triple. Par exemple, dans un VDC avec codage d'effacement par défaut et 4 nœuds, si un nœud tombe en panne, les événements suivants se produisent :

  - La panne du nœud entraîne la perte de 4 fragments.
  - Les fragments manquants sont reconstruits.
  - Le fragment crée 3 copies de réplica, une sur chaque nœud.
  - La copie EC est supprimée.

- Les défaillances un par un agissent comme des défaillances de nœud unique. Par exemple, si vous perdez deux nœuds un par un, chaque défaillance consiste uniquement à récupérer les données de la panne d'un seul nœud.

  Par exemple, avec 6 nœuds et le codage d'effacement par défaut :

  - Première panne : chacun des six nœuds contient maximum trois fragments (16 fragments / 6 nœuds). Les trois fragments manquants sont recréés sur les nœuds restants. Une fois la récupération terminée, le VDC se retrouve avec 5 nœuds.
  - Deuxième panne : chacun des cinq nœuds contient jusqu'à 4 fragments (16 fragments / 5 nœuds). Les quatre fragments manquants sont recréés sur les nœuds restants. Une fois la récupération terminée, le VDC se retrouve avec 4 nœuds.
  - Troisième panne : chacun des quatre nœuds contient jusqu'à 4 fragments (16 fragments / 4 nœuds). Les quatre fragments manquants sont recréés sur les nœuds restants. Une fois la récupération terminée, le VDC se retrouve avec 3 nœuds et, comme ce nombre est inférieur au minimum pour le codage d'effacement, le fragment codé par effacement est remplacé par trois copies de réplica distribuées sur les nœuds restants.
  - Quatrième panne : chacun des trois nœuds contient une copie de réplica. La copie de réplica manquante est recréée sur l'un des nœuds restants. Une fois la récupération terminée, le VDC se retrouve avec 2 nœuds.
  - Cinquième panne : il existe trois copies de réplica, deux sur un nœud et une sur l'autre nœud. Les copies de réplica manquantes sont recréées sur le nœud restant. Une fois la récupération terminée, le VDC se retrouve avec 1 nœud.

# 3 Présentation de la conception multisite

Au-delà de la disponibilité du système et de la durabilité des données conçues au sein d'un site unique, ECS inclut également une protection contre une panne complète à l'échelle du site. Pour se faire, ECS regroupe plusieurs VDC et sites, et configure la géoréplication dans un déploiement multisite.

La fédération de sites implique la fourniture de points de terminaison de réplication et de gestion pour la communication entre les sites. Une fois les sites fédérés, ils peuvent être gérés comme une infrastructure unique.

Les règles de groupe de réplication déterminent la façon dont les données sont protégées et d'où elles sont accessibles. ECS prend en charge à la fois la réplication géoactive et la réplication géopassive. La réplication geoactive fournit un accès actif-actif aux données, ce qui permet de la lire et de l'écrire à partir de n'importe quel site de son groupe de réplication défini.

Lors de la réplication de gamme All-Flash comme  EXF900 sur l'ensemble des sites, il convient de tenir compte des impacts potentiels des performances sur le réseau WAN. L'ingestion volumineuse peut placer une charge élevée sur le lien, entraînant une saturation ou un retard dans la perte de données maximale admissible (RPO). Un utilisateur ou une application peut également subir des temps de latence plus élevés sur les lectures et écritures à distance par rapport aux demandes locales. L'autre doit tenir compte de l'échec partiel du nettoyage de la mémoire, l'ingestion importante à partir du site local et du site de réplication peut accélérer le rythme auquel le système atteint les 90 %, ce qui l'empêchera d'écrire et de récupérer des données et de récupérer.

---

**Remarque :** Nettoyage de la mémoire partielle. Lorsqu'un fragment est à 2/3 nettoyé, ses parties valides sont fusionnées avec d'autres fragments partiellement remplis au sein d'un nouveau fragment, puis l'espace est récupéré.

---

La réplication peut également être géopassive, ce qui signifie que deux à quatre sites sont utilisés comme sources et un ou deux sites sont utilisés uniquement comme cible de réplication. Les cibles de réplication sont utilisées à des fins de récupération uniquement. Les cibles de réplication bloquent l'accès client direct pour les opérations de création, de mise à jour et de suppression.

Les avantages de la réplication géopassive sont les suivants :

- Elle peut optimiser l'efficacité du stockage en augmentant les chances d'exécution des opérations XOR en veillant à ce que les écritures des deux sites sources soient exécutées sur la même cible de réplication.
- Elle permet à l'administrateur de contrôler l'emplacement de la copie de réplication des données, par exemple dans un scénario de sauvegarde vers le Cloud.

ECS offre des options de configuration de géoréplication au niveau du bucket, ce qui permet à l'administrateur de configurer différents niveaux de réplication pour différents buckets.

La Figure 7 montre un exemple de la façon dont un administrateur peut configurer la réplication de trois buckets :

- Bucket A : Données de test/développement d'ingénierie - ne pas répliquer, à conserver localement uniquement
- Bucket B : données de vente européennes - réplication entre les sites en Europe uniquement
- Bucket C : Données de formation à l'échelle de l'entreprise - réplication sur tous les sites de l'entreprise

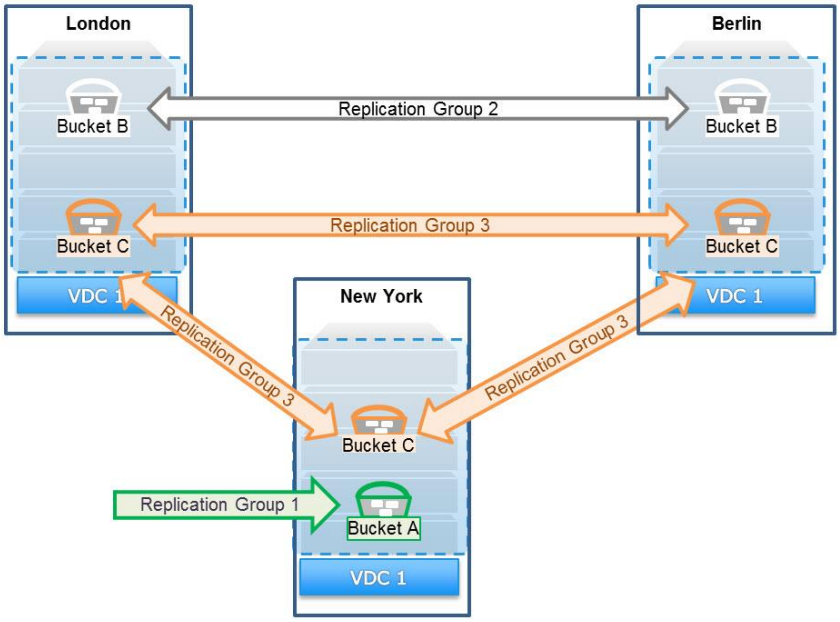


Figure 7 Exemple de différents buckets utilisant différentes règles de réplication

Il est recommandé de configurer des groupes de réplication pour des chemins de réplication spécifiques. Par exemple, il existe un groupe de réplication à la Figure 7 qui réplique les données entre Londres et Berlin. Il devrait être utilisé pour tous les buckets nécessitant la réplication entre Londres et Berlin uniquement.

Les données géorépliquées sont protégées par le stockage d'une copie primaire des données sur le site local et d'une copie secondaire des données sur un ou plusieurs sites distants. Le nombre de copies et l'espace occupé par la copie secondaire sont déterminés en fonction du nombre de sites configurés dans le groupe de réplication, de la façon dont les données sont écrites sur les sites, ainsi que de l'option **Replicate to all site sites**.

Chaque site est responsable de la protection des données locales, ce qui signifie que les copies locales et secondaires protègeront individuellement les données à l'aide du codage d'effacement et/ou de la mise en miroir triple. Le codage d'effacement sur chaque site n'a pas besoin d'être identique sur tous les sites, ce qui signifie qu'un site peut utiliser le codage d'effacement par défaut (12+4) et l'autre site peut utiliser le schéma de codage d'effacement par stockage à froid (10+2).

Les données répliquées sont chiffrées (AES256) et compressées avant d'être envoyées à l'autre site via HTTP.

Afin de maintenir la cohérence entre les sites, il doit y avoir une autorité responsable de la maintenance de la dernière version des métadonnées. L'autorité est définie au niveau du site et détermine la propriété des espaces de nommage, des buckets et des objets. Les informations de propriété sont d'abord stockées sur le site propriétaire, mais sont également répliquées vers les autres sites dans le cadre des métadonnées ECS.

- Version d'autorité : la version faisant autorité est toujours la version propriétaire et est utilisée pour assurer une forte cohérence.
- Version répliquée : la ou les versions répliquées peuvent ne pas être les plus récentes, mais sont utilisées lors des opérations de défaillance, notamment :
  - Si la fonction **Access During Outage** est activée (cohérence éventuelle).
  - Et pendant les opérations de basculement permanentes du site.

Il existe des versions d'autorité du bucket et des propriétaires d'objets.

**Propriétaire de l'espace de nommage :**

- Le site qui crée l'espace de nommage est le propriétaire de l'espace de nommage.
- Il est responsable de la maintenance de la version d'autorité de la liste de buckets.

**Propriétaire du bucket :**

- Le site qui crée le bucket est le propriétaire du bucket.
- Il est responsable de la maintenance de la version d'autorité de :
  - La liste des buckets qui inclut la version la plus récente des objets qui se trouvent dans un bucket.
  - La liste des propriétés d'objets pour les objets au sein de son bucket

**Propriétaire de l'objet :**

- Initialement, le site qui a créé l'objet en premier est le propriétaire de l'objet. Cela peut changer, reportez-vous à la section « Access During Outage » pour plus d'informations.
- Il est responsable de la maintenance de la version d'autorité des métadonnées de l'objet.

## 3.1 Tables de gestionnaire de fragments

Les emplacements des fragments sont conservés dans la table du gestionnaire de fragments, qui est stockée indépendamment sur tous les sites. L'emplacement de création initiale d'un fragment est appelé site principal ; l'emplacement vers lequel il est répliqué est appelé site secondaire.

Lorsqu'un fragment est créé, les sites principal et secondaire sont déterminés et le site principal transmet les informations de site des fragments vers les autres nœuds du groupe de réplication.

De plus, chaque site conservant sa propre table de gestionnaire de fragments, ils contiennent également les informations sur le type de propriété pour chaque fragment. Les types de propriétés sont les suivants :

- **Locale :** sur le site sur lequel le fragment a été créé
- **Copie :** sur le site sur lequel le fragment a été répliqué
- **À distance :** sur le ou les sites sur lesquels ni le fragment ni son réplica ne sont stockés localement
- **Parité :** sur les fragments qui contiennent le résultat d'une opération XOR d'autres fragments (voir la section XOR ci-dessous pour plus de détails)
- **Encodés :** sur les fragments dont les données ont été remplacées localement par des données XOR (voir la section XOR ci-dessous pour plus de détails).

Les Tableau 9 à Tableau 11 montrent des exemples de portions des listes de tables de gestionnaire de fragments de chacun des trois sites.

Tableau 9 Exemple de table de gestionnaire de fragments du Site 1

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Local |
| C2 | Site 2 | Site 3 | Remote |
| C3 | Site 1 | Site 3 | Local |
| C4 | Site 2 | Site 1 | Copy |

Tableau 10 Exemple de table de gestionnaire de fragments du Site 2

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Copy |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 1 | Site 3 | Remote |
| C4 | Site 2 | Site 1 | Local |

Tableau 11 Exemple de table de gestionnaire de fragments du Site 3

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Remote |
| C2 | Site 2 | Site 3 | Copy |
| C3 | Site 1 | Site 3 | Copy |
| C4 | Site 2 | Site 1 | Remote |

## 3.2 Codage XOR

Afin d'optimiser l'efficacité du stockage des données configurées avec un groupe de réplication contenant au moins trois sites, ECS utilise le codage XOR. Plus le nombre de sites dans un groupe de réplication augmente, plus l'algorithme ECS est efficace pour réduire les surcharges.

Le codage XOR est effectué sur chaque site. Il analyse sa table de gestionnaire de fragments et, lorsqu'il trouve des fragments de type COPY provenant de chacun des autres sites de son groupe de réplication, il peut effectuer un codage XOR sur ces fragments. Par exemple, dans la Table 12, il présente le Site 3 d'une configuration à trois sites avec les fragments **C2** et **C3** qui sont de type COPY chacun avec un site principal différent. Cela permet au Site 3 de les intégrer au XOR et de stocker le résultat. Le résultat est un nouveau fragment, **C5** qui est un XOR de **C2** et **C3** (mathématiquement C2 $\oplus$ C3) et il est répertorié comme fichier de **Parité**, sans site secondaire. Les ID de fragment des fragments de parité ne sont pas transmis à d'autres sites.

La Tableau 12 montre un exemple de table de gestionnaire de fragments sur le Site 3 alors qu'il exécute XOR des fragments **C2** et **C3** ensemble pour créer le fragment **C5**.

Tableau 12 Table du gestionnaire de fragments du Site 3 pendant l'opération XOR

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Remote |
| C2 | Site 2 | Site 3 | Copy |
| C3 | Site 1 | Site 3 | Copy |
| C4 | Site 2 | Site 1 | Remote |
| C5 | Site 3 | | Parité (C2 et C3) |

Une fois le codage XOR terminé, la copie des données pour **C2** et **C3** est supprimée, ce qui libère de l'espace sur le disque, et le type de fichier dans la table du gestionnaire de fragments pour ces fragments passe à Encoded. L'opération XOR est uniquement une opération de site secondaire, le site principal ne sait pas que ses fragments ont été codés. Une fois le codage XOR terminé et la copie des données pour **C2** et **C3** supprimée, la table du gestionnaire de fragments du Site 3 est répertorié comme illustré à la Tableau 13.

Tableau 13 Table du gestionnaire de fragments du Site 3 après avoir terminé le codage XOR de C2 et C3

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Remote |
| C2 | Site 2 | Site 3 | Codé |
| C3 | Site 1 | Site 3 | Codé |
| C4 | Site 2 | Site 1 | Remote |
| C5 | Site 3 | | Parité (C2 et C3) |

Les demandes de données dans un fragment codé seront traitées par le site contenant la copie primaire. Pour plus d'informations sur l'efficacité du stockage, reportez-vous au livre Blanc : ECS Overview and Architecture.

## 3.3 Option de réplication sur tous les sites Replicate to all sites

Replicate to all sites est une option de groupe de réplication qui est utilisée lorsque vous disposez de trois sites ou plus et que vous souhaitez répliquer tous les fragments sur tous les VDC ou sites configurés dans le groupe de réplication. Cela empêche également l'exécution des opérations XOR. Lorsque l'option **Replicate to all sites** est :

- **Activée :** le nombre de copies de données écrites est égal au nombre de sites dans le groupe de réplication. Par exemple, si vous avez quatre sites dans le groupe de réplication, vous aurez une copie primaire plus trois copies secondaires : une sur chaque site.
- **Désactivée :** le nombre de copies de données écrites est de deux. Vous disposez de la copie primaire et d'une copie répliquée sur un site distant, quel que soit le nombre total de sites.

**Remarque :** L'option Replicate to all sites ne peut pas être activée pour un groupe de réplication avec une configuration géopassive.

Ce paramètre n'a aucun impact sur les groupes de réplication qui ne contiennent que deux sites puisque les données sont déjà toutes répliquées sur les deux sites. L'administrateur peut choisir les buckets qui utilisent ce groupe de réplication.

L'activation de l'option Replicate to all sites a les effets suivants :

- Elle peut améliorer les performances de lecture, car une fois la réplication terminée, les lectures suivantes sont gérées localement.
- Elle supprime l'impact sur les performances causé par le décodage XOR.
- Elle offre une durabilité des données accrue.
- Elle réduit l'efficacité du taux d'utilisation du stockage.
- Elle augmente l'utilisation du réseau WAN pour la géoréplication ; qui est proportionnelle au nombre de VDC ou de sites dans le groupe de réplication.

- Elle réduit l'utilisation du réseau WAN pour les lectures de données répliquées.

Pour ces raisons, l'activation de cette option est uniquement recommandée pour des buckets spécifiques dans des environnements qui remplissent les critères suivants :

- Les environnements dont la charge applicative de lecture est gourmande pour les mêmes données provenant de sites géographiquement dispersés.
- Les environnements dont l'infrastructure dispose d'une bande passante WAN suffisante entre les sites du groupe de réplication pour permettre l'augmentation du trafic de géoréplication.
- Les environnements qui privilégient les performances de lecture au détriment de l'efficacité du taux d'utilisation du stockage.

## 3.4 Écriture d'un flux de données dans un environnement géorépliqué

Les fragments contiennent 128 Mo de données d'un ou plusieurs objets de buckets qui partagent les mêmes paramètres de groupe de réplication. La réplication n'est pas effectuée simultanément, et elle est lancée au niveau du fragment par le propriétaire de la partition de fragment. Si le fragment est configuré avec géoréplication, les données sont ajoutées à une file d'attente de réplication au fur et à mesure qu'elles sont écrites dans le fragment du site principal, il n'attend pas que le fragment soit verrouillé. Des opérateurs de threads d'E/S traitent en continu la file d'attente.

L'opération d'écriture se produit d'abord localement, y compris l'ajout de la protection des données, puis elle est répliquée et protégée sur le site distant. La Figure 8 montre un exemple du processus d'écriture d'un objet de 128 Mo vers un bucket géorépliqué.

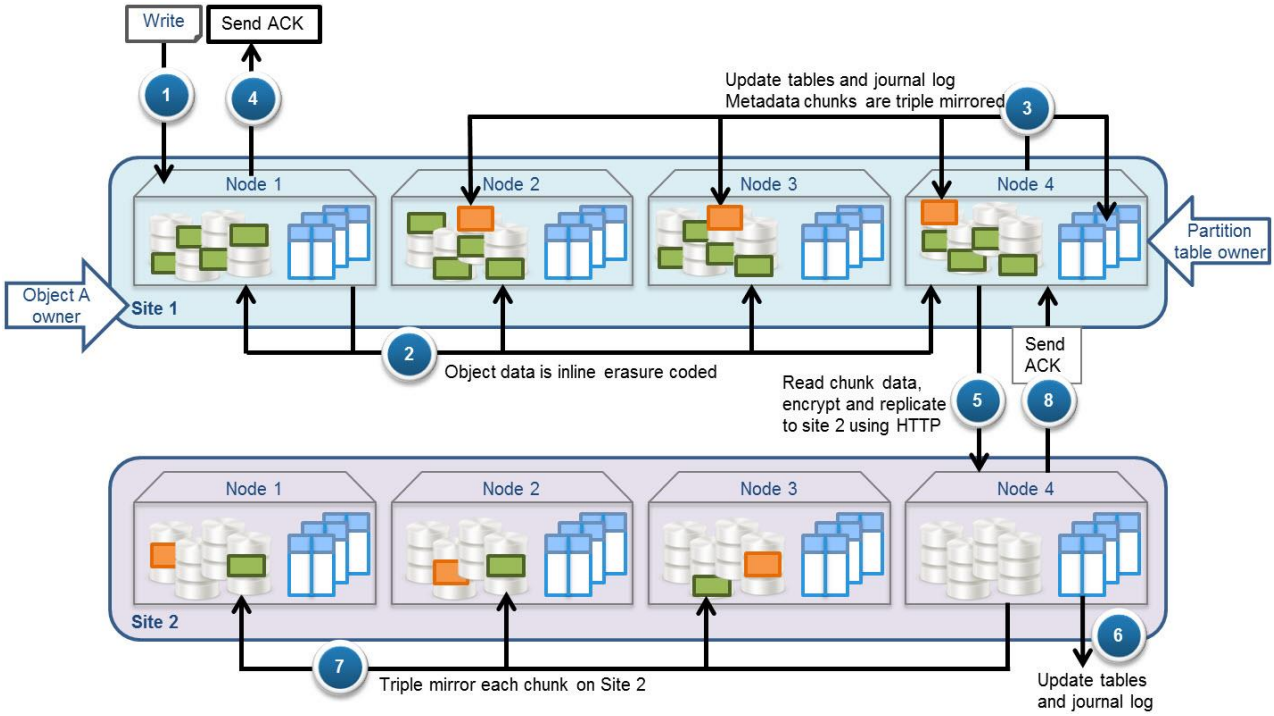


Figure 8 Workflow d'écriture de données d'un objet de 128 Mo vers un bucket

1. La demande d'écriture de l'Objet A est envoyée à un nœud, dans cet exemple, Site 1 Nœud 1. Le Site 1 devient le propriétaire de l'Objet A.
2. Les données sont codées par effacement dans la ligne et écrites dans un fragment du Site 1.
3. Les propriétaires de partitions de table, dans cet exemple, Nœud 4, mettent à jour les tables concernées (par exemple, les tables de listes de fragments, d'objets et de buckets) et enregistrent les transactions dans les journaux log. Ces métadonnées sont écrites dans un fragment de métadonnées qui est mis en miroir triple sur le Site 1.

4. Un accusé de réception confirmant l'écriture est envoyé au client.
5. Pour chaque fragment, le propriétaire de la table de partition de fragments, dans cet exemple, Nœud 4 :
   a. ajoute les données dans le fragment à la file d'attente de réplication une fois qu'elles sont écrites localement sans attendre que le fragment soit verrouillé.
   b. lit les éléments de données du fragment (les fragments de parité ne sont lus que si nécessaire pour recréer un fragment de données manquant).
   c. chiffre et réplique les données sur le Site 2 via HTTP.
6. Les propriétaires de partitions de table pour les fragments répliqués, dans cet exemple, Site 2 Nœud 4, mettent à jour les tables concernées et enregistrent les transactions dans les journaux log qui sont mis en miroir triple.
7. Chaque fragment répliqué est d'abord écrit sur le deuxième site avec une mise en miroir triple.
8. Un accusé de réception est renvoyé au propriétaire de la table de partition de fragments du site principal.

**Remarque :** Les données écrites sur le site répliqué seront codées par effacement à l'issue d'un délai, ce qui laisse le temps à d'autres processus, tels que les opérations XOR, de se terminer en premier.

## 3.5 Lecture d'un flux de données dans un environnement géorépliqué

Puisqu'ECS ne réplique pas les données simultanément sur plusieurs VDC au sein d'un groupe de réplication, il nécessite une méthode pour garantir la cohérence des données entre les sites et les VDC. ECS garantit une forte cohérence en récupérant la dernière copie des métadonnées à partir du site propriétaire de l'objet. Si le site demandeur contient une copie (type de fragment = local ou copie) de l'objet, il l'utilisera pour répondre de la demande de lecture, sinon il récupérera les données auprès du propriétaire de l'objet. Un exemple de lecture de flux de données est illustré à la Figure 9qui décrit une demande de lecture de l'Objet A d'un site autre que le site propriétaire de l'objet.

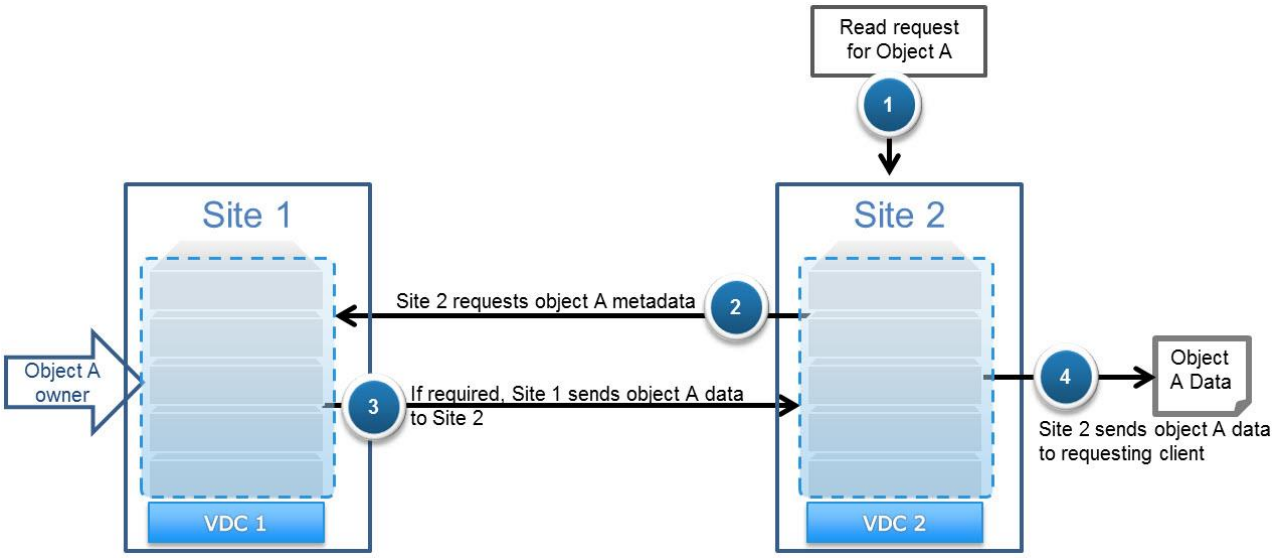


Figure 9 Workflow de lecture de données d'un objet géorépliqué appartenant à un autre site.

Dans cet exemple, le flux de l'opération de lecture s'affiche comme suit :

1. Le Site 2 reçoit une demande de lecture pour l'Objet A qui appartient au Site 1.

2. Le Site 2 contacte le propriétaire du bucket de l'objet, dans cet exemple Site 1, pour obtenir la dernière version des métadonnées.
   Propriété de l'objet :
   - Si la fonction **Access During Outage** est désactivée, il vérifie ses informations locales pour déterminer s'il s'agit du propriétaire de l'objet. Si ce n'est pas le cas, il contactera le propriétaire du bucket pour déterminer qui est le propriétaire de l'objet.
   - Si la fonction **Access During Outage** est activée sur le bucket, le site demandeur vérifie auprès du propriétaire du bucket qui est le propriétaire actuel de l'objet.
3. Si le Site 2 ne dispose pas de copie de l'objet (type de fragment = local ou copie), le Site 1 envoie alors les données de l'Objet A au Site 2.
4. Le Site 2 envoie les données de l'Objet A au client demandeur.

## 3.6 Mise à jour d'un flux de données dans un environnement géorépliqué

ECS est conçu pour permettre la mise à jour active-active des données des nœuds au sein du groupe de réplication des buckets associés. Pour ce faire, les sites non-propriétaires d'objets doivent envoyer simultanément des informations sur une mise à jour d'objet au site propriétaire primaire et attendre l'accusé de réception avant de pouvoir renvoyer l'accusé de réception au client. Les données de l'objet mis à jour sont répliquées dans le cadre des activités normales de réplication asynchrone des fragments. Si les données ne sont pas encore répliquées sur le site propriétaire et qu'il reçoit une demande de lecture pour les données, il demandera les données au site distant. La Figure 10 montre un exemple illustrant une demande de mise à jour de l'Objet A d'un site autre que le site propriétaire de l'objet.

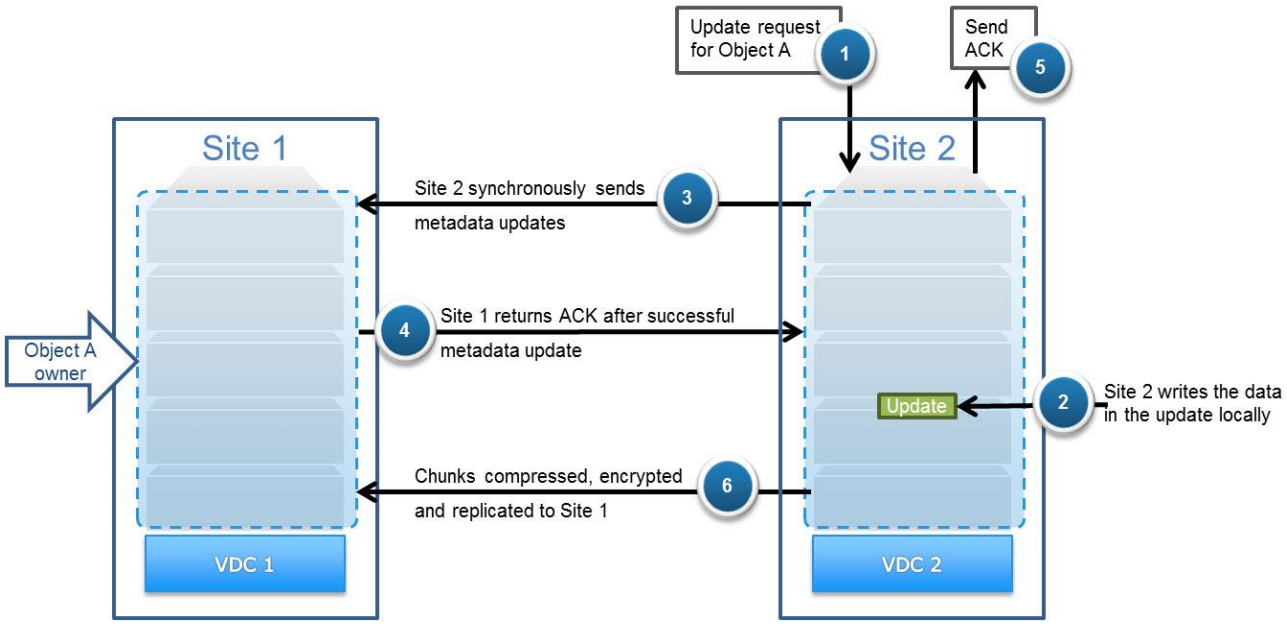


Figure 10 Workflow de la mise à jour pour un objet géorépliqué appartenant à un autre site.

Dans cet exemple, le flux de l'opération de mise à jour se déroule comme suit :

1. Le Site 2 reçoit une demande de mise à jour pour l'Objet A qui appartient au Site 1.
2. Le Site 2 écrit localement les données dans la mise à jour.
3. Le Site 2 envoie les mises à jour de métadonnées simultanément au propriétaire de l'objet, dans cet exemple Site 1.

**Remarque :** Si la fonction **Access During Outage** est activée sur le bucket, ou si le Site 2 n'est pas le propriétaire de l'objet, il contacte d'abord le propriétaire du bucket pour déterminer qui est le propriétaire actuel de l'objet.

4. Le Site 1 envoie un accusé de réception au Site 2 indiquant que la mise à jour des métadonnées a réussi.
5. Le Site 2 envoie l'accusé de réception au client demandeur indiquant que la mise à jour a réussi.
6. Les fragments sont ajoutés à la file d'attente de réplication, chiffrés et répliqués de manière asynchrone sur le Site 1 comme d'habitude.

Normalement, le propriétaire de l'objet ne change pas après une mise à jour, quel que soit le site d'où provient la mise à jour. Dans cet exemple, le Site 1 reste le propriétaire de l'objet, même après la mise à jour réussie originaire du Site 2. La seule exception concerne les pannes de site temporaires si la fonction **Access During Outage** est activée. Reportez-vous à la section 4.1.2 pour plus de détails.

# 4 Disponibilité multisite

ECS offre une forte cohérence, et exige que les demandes d'E/S soient vérifiées avec le propriétaire avant de répondre. Pour cette raison, si un site est inaccessible, l'accès à certains buckets et objets peut être temporairement interrompu.

La durée et les raisons des pannes de site peuvent varier :

- Elles peuvent être temporaires et dues, par exemple, à la perte de la connectivité réseau entre les sites fédérés ou à la défaillance d'un site entier, comme une panne d'alimentation du bâtiment.
- Elles peuvent être permanentes en cas de catastrophe naturelle, par exemple.

Pour détecter les pannes de site temporaires, les sites fédérés émettent des pulsations entre les sites. Si la pulsation est perdue entre les sites pendant une période continue, la valeur par défaut étant de 15 minutes :

- Chaque site indique l'autre comme étant en échec, dans une configuration à deux sites.
- Dans une configuration à trois sites ou plus, un site ne sera marqué comme étant en panne que si les deux conditions suivantes sont remplies :
  - La majorité des sites perdent des pulsations pendant la période continue vers le même site ECS
  - Tous les sites restants sont actuellement marqués comme étant en ligne

Par exemple, dans une configuration à trois sites, si les sites 2 et 3 perdent la connectivité réseau au Site 1 pendant une période continue, ECS marque le Site 1 comme étant temporairement en panne.

En cas d'échec d'un site fédéré, la disponibilité du système peut être maintenue en redirigeant l'accès à d'autres systèmes fédérés. Au cours de la panne à l'échelle du site, les données géorépliquées détenues par le site indisponible deviennent temporairement indisponibles. La durée pendant laquelle les données restent indisponibles est déterminée par les éléments suivants :

- Si la fonction **Access During Outage** est activée ou non
- Durée de la panne temporaire du site
- Temps nécessaire au basculement d'un site permanent pour effectuer les opérations de récupération

La panne du site peut être temporaire ou permanente. Une panne temporaire du site signifie que le site peut être remis en ligne. Elle est généralement causée par des pannes d'alimentation ou une perte de gestion de réseau entre les sites. Une panne de site permanente se produit lorsque l'ensemble du système est irrécupérable, par exemple à cause d'un incendie en laboratoire. Seul un administrateur peut déterminer si une panne de site est permanente et si des opérations de récupération sont nécessaires.

## 4.1 Panne de site temporaire (TSO)

Des pannes de site temporaires se produisent lorsqu'un site est temporairement inaccessible aux autres sites du groupe de réplication. ECS permet aux administrateurs de disposer de deux options de configuration qui déterminent la façon dont les objets sont accessibles lors d'une panne de site temporaire.

- Désactivez l'option **Access During Outage** (ADO) qui conserve une forte cohérence en :
  - Continuant à autoriser l'accès aux données détenues par les sites accessibles.
  - Empêchant l'accès aux données détenues par un site inaccessible.
- Activez l'option **Access During Outage** qui autorise l'accès en lecture et, éventuellement, en écriture à toutes les données géorépliquées, y compris celles détenues par le site signalé comme étant en échec. Au cours d'une TSO avec l'option **Access During Outage** activée, les données du bucket basculent temporairement sur la cohérence finale ; lorsque tous les sites seront de nouveau en ligne, celles-ci bénéficieront à nouveau d'un niveau de cohérence élevé.

L'option **Access During Outage** est désactivée par défaut.

L'option **Access During Outage** peut être définie au niveau du bucket ; cela signifie que vous pouvez activer cette option pour certains buckets et pas pour d'autres. Cette option de bucket peut être modifiée à tout moment tant que tous les sites sont en ligne, ils ne peuvent pas être modifiés en cas de panne du site.

Lors d'une panne de site temporaire :

- Les buckets, les espaces de nommage, les utilisateurs d'objets, les fournisseurs d'authentification, les groupes de réplication et les mappages d'utilisateurs et de groupes NFS ne peuvent pas être créés, supprimés ou mis à jour depuis n'importe quel site (les groupes de réplication peuvent être supprimés d'un VDC lors d'un basculement de site permanent).
- Vous ne pouvez pas répertorier les buckets d'un espace de nommage lorsque le site propriétaire de l'espace de nommage n'est pas accessible.
- Les systèmes de fichiers au sein de buckets HDFS/NFS sont détenus par le site non disponible sont en lecture seule.
- Lorsque vous copiez un objet à partir d'un bucket détenu par le site non disponible, la copie est une copie complète de l'objet source. Cela signifie que les mêmes données d'objet sont stockées plusieurs fois. Dans des conditions normales, la copie de l'objet se compose des indices de données de l'objet, et pas d'une copie complète des données de l'objet.
- Les utilisateurs OpenStack Swift ne peuvent pas se connecter à OpenStack pendant une TSO, car ECS ne peut pas authentifier les utilisateurs Swift au cours de la TSO. Après la TSO, les utilisateurs Swift doivent se réauthentifier.

### 4.1.1 Comportement TSO par défaut

Étant donné qu'ECS offre une forte cohérence, les demandes d'E/S nécessitent une vérification auprès du propriétaire avant de répondre. Si un site est inaccessible à d'autres sites au sein d'un groupe de réplication, certains accès aux buckets et aux objets peuvent être perturbés.

La Tableau 14 indique l'accès requis pour qu'une opération réussisse.

Tableau 14 Exigences d'accès

| Operation | Conditions de réussite |
|---|---|
| Création d'un objet | Nécessite que le propriétaire du bucket soit accessible |
| Liste des objets | Nécessite que le propriétaire du bucket et tous les objets du bucket soient accessibles au nœud demandeur |
| Lecture d'un objet<br><br>Mise à jour d'un objet | Nécessite que :<br>• Le propriétaire de l'objet et du bucket (la propriété du bucket n'est requise que si l'option **Access During Outage** est activée sur le bucket contenant l'objet)<br>• Ou que le propriétaire de l'objet et le propriétaire du bucket soient accessibles par le nœud demandeur |

- Une opération de création d'objets inclut la mise à jour de la liste des buckets avec le nouveau nom d'objet. Cela nécessite l'accès au propriétaire du bucket et, par conséquent, échoue si le site demandeur n'a pas accès au propriétaire du bucket.

- La création d'une liste d'objets dans un bucket nécessite à la fois la création d'une liste d'informations provenant du propriétaire du bucket et les informations en en-tête de chaque objet du bucket. Par conséquent, les demandes de liste de buckets suivantes échoueront si l'option **Access During Outage** est désactivée :
  - Demandes de répertorier les buckets appartenant à un site qui n'est pas accessible au demandeur.
  - Bucket contenant des objets appartenant à un site qui n'est pas accessible au demandeur.
- La lecture de l'objet nécessite tout d'abord de lire les métadonnées de l'objet du propriétaire de l'objet.
  - Si le site demandeur est le propriétaire de l'objet et que l'option **Access During Outage** est désactivée, la demande aboutira.
  - Si le site demandeur est le propriétaire de l'objet et du bucket, la demande aboutira.
  - Si le propriétaire de l'objet n'est pas local, le site doit vérifier auprès du propriétaire du bucket pour trouver le propriétaire de l'objet. Si le propriétaire de l'objet ou les sites propriétaires du bucket ne sont pas disponibles pour le demandeur, l'opération de lecture échouera.
- La mise à jour de l'objet nécessite une mise à jour complète des métadonnées sur le propriétaire de l'objet.
  - Si le site demandeur est le propriétaire de l'objet et que l'option **Access During Outage** est désactivée, la demande aboutira.
  - Si le site demandeur est le propriétaire de l'objet et du bucket, la demande aboutira.
  - Si le propriétaire de l'objet n'est pas local, le site doit vérifier auprès du propriétaire du bucket pour trouver le propriétaire de l'objet. Si le propriétaire de l'objet ou les sites propriétaires du bucket ne sont pas disponibles pour le demandeur, l'opération de lecture échouera.

La Figure 11 montre la disposition du bucket et de l'objet si l'on prend un exemple à trois sites.

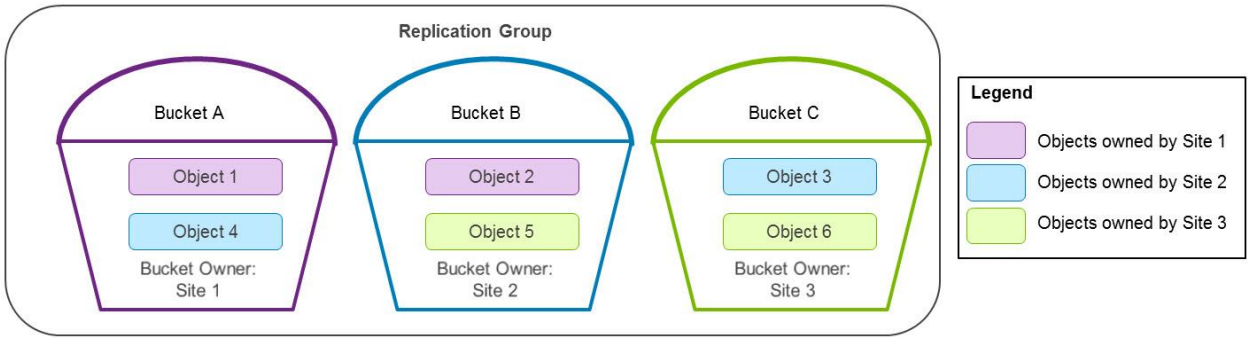


Figure 11 Exemple de propriété de bucket et d'objet

La Tableau 15 répertorie les opérations qui réussiront ou échoueront dans l'exemple de configuration à trois sites, illustré à la Figure 11 si le Site 1 est inaccessible aux autres sites du groupe de réplication. Pour simplifier la lecture de la table, le site inaccessible est répertorié comme étant en échec et les deux autres sont répertoriés en ligne.

Tableau 15 Opérations qui réussissent ou échouent si le Site 1 est inaccessible aux autres sites

| Operation | Bucket / objet | Demande envoyée à | | |
|---|---|---|---|---|
| | | **Site 1 (échec)** | **Site 2 (en ligne)** | **Site 3 (en ligne)** |
| Création d'un objet dans | Bucket A | Success<br>Bucket détenu localement | Échec<br>Impossible d'accéder au propriétaire du bucket | Échec<br>Impossible d'accéder au propriétaire du bucket |
| | Bucket B | Échec<br>Impossible d'accéder au propriétaire du bucket | Success<br>Bucket détenu localement | Success<br>Bucket détenu par un site en ligne |
| | Bucket C | Échec<br>Impossible d'accéder au propriétaire du bucket | Success<br>Bucket détenu par un site en ligne | Success<br>Bucket détenu localement |
| Liste des objets dans | Bucket A | Échec<br>Bien que le bucket soit détenu localement, il contient un objet appartenant à un site auquel il ne peut pas accéder | Échec<br>Impossible d'accéder au propriétaire du bucket | Échec<br>Impossible d'accéder au propriétaire du bucket |
| | Bucket B | Échec<br>Impossible d'accéder au propriétaire du bucket | Échec<br>Bien que le bucket soit détenu localement, il contient un objet appartenant au site en panne | Échec<br>Bien que le propriétaire du bucket soit en ligne, le bucket contient un objet appartenant au site en panne |
| | Bucket C | Échec<br>Impossible d'accéder au propriétaire du bucket | Success<br>Le propriétaire du bucket est un site en ligne et tous les objets proviennent de sites en ligne | Success<br>Bucket détenu localement et tous les objets proviennent de sites en ligne |
| Lecture ou mise à jour d'un objet | Objet 1 | Success<br>Objet détenu localement | Échec<br>Impossible d'accéder au propriétaire de l'objet | Échec<br>Impossible d'accéder au propriétaire de l'objet |
| | Objet 2 | Success<br>Objet détenu localement | Échec<br>Impossible d'accéder au propriétaire de l'objet | Échec<br>Impossible d'accéder au propriétaire de l'objet |
| | Objet 3 | Échec<br>Impossible d'accéder au propriétaire de l'objet | Success<br>Objet détenu localement | Success<br>L'objet n'est pas détenu localement, on retrouve le propriétaire de l'objet depuis le propriétaire du bucket qui est en ligne |
| | Objet 4 | Échec<br>Impossible d'accéder au propriétaire de l'objet | Success<br>Objet détenu localement | Échec<br>L'objet n'est pas détenu localement, obligation d'accéder au propriétaire du bucket qui est le site en panne |

| Operation | Bucket / objet | Demande envoyée à | | |
|---|---|---|---|---|
| | | **Site 1 (échec)** | **Site 2 (en ligne)** | **Site 3 (en ligne)** |
| | Objet 5 | Échec<br>Impossible d'accéder au propriétaire de l'objet | Success<br>L'objet n'est détenu localement, on retrouve le propriétaire de l'objet depuis le propriétaire du bucket qui est le propriétaire de l'objet | Success<br>Objet détenu localement |
| | Objet 6 | Échec<br>Impossible d'accéder au propriétaire de l'objet | Success<br>L'objet n'est pas détenu localement, on retrouve le propriétaire de l'objet depuis le propriétaire du bucket qui est en ligne | Success<br>Objet détenu localement |

### 4.1.2 Comportement TSO avec l'option Access During Outage activée

Lorsqu'un site est inaccessible pour la première fois à d'autres sites au sein d'un groupe de réplication, le comportement est détaillé dans la section comportement TSO par défaut. Quand la pulsation est perdue entre les sites pendant une période continue, la valeur par défaut étant de 15 minutes, l'ECS signale le site comme en panne. L'activation de l'option **Access During Outage** (ADO) sur un bucket modifie le comportement TSO après qu'un site est signal comme étant en panne, ce qui permet aux objets de ce bucket d'utiliser la cohérence éventuelle. Cela signifie qu'après l'échec temporaire d'un site, tous les buckets dont l'option **Access During Outage** est activée prennent en charge les lectures et, éventuellement, les écritures d'un site non-propriétaire. Pour ce faire, il autorise l'utilisation des métadonnées répliquées lorsque la copie d'autorité sur le site propriétaire n'est pas disponible. Vous pouvez modifier l'option **Access During Outage** du bucket à tout moment, sauf lors d'une panne de site.

L'option **Access During Outage** permet d'accéder aux données quand un site est signalé comme étant en panne. L'inconvénient est que les données renvoyées peuvent être obsolètes.

À partir de la version 3.1, une option de bucket supplémentaire pour l'option **read-only access during outage** a été ajoutée. Elle garantit que la propriété de l'objet n'est jamais modifiée et supprime le risque de conflits causés par les mises à jour d'objets sur les sites en panne et les sites en ligne lors d'une panne de site temporaire. L'inconvénient de l'option **read-only access during outage** est que, lors d'une panne de site temporaire, aucun nouvel objet ne peut être créé et aucun objet existant dans le bucket ne peut être mis à jour avant que tous les sites ne soient de nouveau en ligne. L'option **read-only access during outage** est uniquement disponible lors de la création du bucket. Elle ne peut pas être modifiée par la suite.

Comme mentionné précédemment, un site est en panne quand la pulsation est perdue entre les sites pendant une période continue, la valeur par défaut étant de 15 minutes. Par conséquent, si la pulsation est perdue pendant une période continue :

- Dans une configuration à deux sites, chaque site se considérera comme en ligne, et indiquera l'autre comme étant en échec.
- Dans une configuration à trois sites ou plus, un site ne sera marqué comme étant en échec que si les deux :
  - La majorité des sites perdent des pulsations pendant la période continue vers le même site ECS
  - Tous les sites restants sont actuellement marqués comme étant en ligne

Un site en panne peut toujours être accessible par les clients et les applications, par exemple lorsque le réseau interne d'une entreprise perd la connectivité à un site unique, mais que le réseau extranet reste opérationnel. Par exemple, dans une configuration à cinq sites, si les sites 2 à 5 perdent la connectivité réseau au Site 1 pendant une période continue, ECS marque le Site 1 comme étant temporairement en panne. Si le Site 1 est toujours accessible aux clients et aux applications, il sera en mesure de traiter les demandes de service pour les buckets et objets détenus localement, car les recherches sur d'autres sites ne sont pas nécessaires. Toutefois, les demandes adressées au Site 1 pour des buckets et des objets non-détenus échoueront. La Tableau 16 indique accès requis quand un site a été marqué comme étant en panne afin qu'une opération réussisse, si l'option **Access During Outage** est activée.

Tableau 16 Opérations réussies après l'échec d'un site avec l'option Access During Outage activée

| **Operation** | **Demande envoyée au site en panne (dans une fédération de trois sites ou plus)** | **Demande envoyée à un site en ligne, y compris :**<br>• **Tous les sites en ligne dans une fédération de trois sites ou plus**<br>• **Ou sur tous les sites dans une fédération de deux sites** |
|---|---|---|
| Création d'un objet | Réussite pour les buckets détenus localement, sauf si l'option **read-only access during outage** est activée sur le bucket.<br><br>Échec pour les buckets détenus à distance | Réussite, sauf si l'option **read-only access during outage** est activée sur le bucket. |
| Liste des objets | Répertorie uniquement les objets de ses buckets détenus localement si tous les objets sont également détenus localement | Réussite<br>N'inclut pas les objets détenus par un site en panne qui n'ont pas fini d'être répliqués |
| Lecture d'un objet | Réussite pour les objets détenus localement dans des buckets détenus localement<br>(il peut ne pas s'agir de la version la plus récente)<br><br>Échec pour les buckets détenus à distance | Réussite<br>Si l'objet est détenu par le site en panne, il faut que l'objet d'origine ait terminé la réplication avant que la panne ne se produise. |
| Mise à jour d'un objet | Réussite pour les objets détenus localement dans des buckets détenus localement, sauf si l'option **read-only access during outage** est activée sur le bucket.<br><br>Échec pour les buckets détenus à distance | Réussite, sauf si l'option **read-only access during outage** est activée sur le bucket.<br>Acquiert la propriété de l'objet |

- Création d'un objet

  Lorsqu'un site est signalé comme étant en panne, la création d'objets échoue si l'option **read-only access during outage** est activée sur le bucket. Si elle est désactivée :

  - Dans une fédération de trois sites ou plus, le site signalé comme étant en panne, s'il est accessible par les clients ou les applications, peut créer des objets uniquement dans ses buckets détenus localement. Ces nouveaux objets ne sont accessibles qu'à partir de ce site. Les autres sites ne seront pas informés de ces objets tant que le site en échec n'aura pas été remis en ligne et qu'ils auront accès au propriétaire du bucket.

  - Les sites en ligne peuvent créer des objets dans n'importe quel bucket, y compris les buckets détenus par le site signalé comme étant en panne. La création d'un objet nécessite la mise à jour de la liste de buckets avec le nouveau nom d'objet. Si le propriétaire du bucket est indisponible, il crée un historique d'objets qui intègre l'objet dans la table de listes de buckets lors des opérations de récupération ou de réintégration du propriétaire du bucket indisponible.

- Liste d'objet
  - Dans une fédération de trois sites ou plus, le site signalé comme étant en panne nécessite la propriété locale du bucket et de tous les objets du bucket pour répertorier correctement les objets d'un bucket. La liste du site en panne n'inclut pas les objets créés à distance alors qu'il est temporairement en panne.
  - Les sites en ligne peuvent créer des objets dans n'importe quel bucket, y compris dans un bucket détenu par le site panne. Il répertorie la dernière version du bucket qu'il possède. Elle peut être légèrement obsolète.
- Lecture d'un objet
  - Dans une fédération de trois sites ou plus, le site en panne, s'il est accessible par les clients ou les applications, peut lire uniquement des objets détenus localement dans des buckets détenus localement.
  - La demande de lecture sur un site en panne doit d'abord avoir accès au propriétaire du bucket pour valider le propriétaire actuel de l'objet. S'il peut accéder au propriétaire du bucket et que le propriétaire actuel de l'objet est local, la demande de lecture aboutira. Si le propriétaire du bucket ou le propriétaire actuel de l'objet n'est pas accessible, la demande de lecture échouera.
  - Les sites en ligne peuvent lire tous les objets, y compris ceux appartenant au site en panne tant que la réplication de l'objet d'origine est terminée. Il vérifie l'historique des objets et fournit la dernière version de l'objet disponible. Si un objet a été mis à jour par la suite sur le site en panne et que la géoréplication de la version mise à jour ne s'est pas terminée, l'ancienne version sera utilisée pour répondre à la demande de lecture.

**Remarque :** les demandes de lecture envoyées aux sites en ligne où le propriétaire du bucket est le site en panne utiliseront ses informations sur la liste des buckets locaux et son historique des objets pour déterminer le propriétaire de l'objet.

- Mise à jour d'un objet
  - Lorsqu'un site est signalé comme étant en panne, la mise à jour d'objets échoue si l'option **read-only access during outage** est activée sur le bucket. Dans une fédération de trois sites ou plus, le site en panne, s'il est accessible par les clients ou les applications, peut mettre à jour uniquement des objets détenus localement dans des buckets détenus localement.
  - La demande de mise à jour doit d'abord avoir accès au propriétaire du bucket pour valider la propriété actuelle de l'objet. S'il peut accéder au propriétaire du bucket et que le propriétaire actuel de l'objet est local, la demande de mise à jour aboutira. Si le propriétaire du bucket ou le propriétaire actuel de l'objet n'est pas accessible, la demande de mise à jour échouera.
  - Une fois les opérations de réintégration terminées, cette mise à jour ne sera pas incluse dans les opérations de lecture si le site distant a également mis à jour le même objet au cours de la même TSO.

  Un site en ligne peut mettre à jour les objets détenus par les sites en ligne et les sites en panne. Si une demande de mise à jour d'objet est envoyée à un site en ligne pour un objet détenu par le site en panne, elle met à jour la dernière version de l'objet disponible sur un système en ligne.

  Le site qui effectue la mise à jour devient le nouveau propriétaire de l'objet et met à jour l'historique des objets avec les informations du nouveau propriétaire et le nouveau numéro de séquence. Il sera utilisé lors des opérations de récupération ou de réintégration du propriétaire d'objet d'origine pour mettre à jour son historique d'objets avec le nouveau propriétaire.

**Remarque :** Les demandes de mise à jour envoyées aux sites en ligne où le propriétaire du bucket est le site en panne utiliseront ses informations sur la liste des buckets locaux et son historique des objets pour déterminer le propriétaire de l'objet.

Cet exemple montre le scénario avec une disposition bucket et objet pour l'espace de nommage 1 dans une configuration à trois sites, comme illustré à la Figure 12.

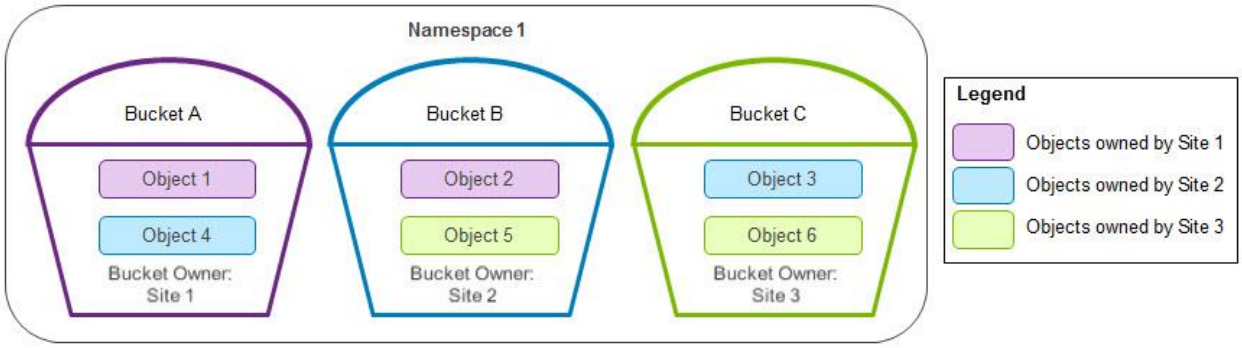


Figure 12 Propriété des buckets et des objets pour l'espace de nommage 1

Dans cette configuration à trois sites, la Tableau 17 montre un exemple si les trois conditions suivantes sont vraies :

- L'option **Access During Outage** est activée
- Et l'option read-only access during outage est désactivée
- Et le Site 1 est en panne

Tableau 17 Exemple d'opérations qui réussissent ou échouent avec les options **Access During Outage** et **read-only access during outage** désactivées, avec le Site 1 temporairement en panne dans une configuration à trois sites

| Operation | Bucket / objet | Demande envoyée à | |
|---|---|---|---|
| | | **Site 1 (échec)** | **Site 2 ou Site 3 (en ligne)** |
| Création d'un objet dans | Bucket A | Success | Success |
| | Bucket B | Échec<br>Le site en panne peut uniquement créer des objets dans des buckets détenus localement | Success |
| | Bucket C | Échec<br>Le site en panne peut uniquement créer des objets dans des buckets détenus localement | Success |
| Liste des objets dans | Bucket A | Échec<br>Bien que le bucket soit détenu localement, il contient des objets détenus à distance | Success<br>N'inclut pas les objets détenus par un site en panne qui n'ont pas été répliqués |
| | Bucket B | Échec<br>Le site en panne peut uniquement répertorier des objets dans des buckets détenus localement | Success |
| | Bucket C | Échec<br>Le site en panne peut uniquement répertorier des objets dans des buckets détenus localement | Success |

| Operation | Bucket / objet | Demande envoyée à | |
|---|---|---|---|
| | | Site 1 (échec) | Site 2 ou Site 3 (en ligne) |
| Lecture ou mise à jour d'un objet | Objet 1 | Réussite : les objets et les buckets sont détenus localement | Success<br>L'objet doit avoir terminé sa réplication avant le TSO.<br>La mise à jour acquiert la propriété de l'objet |
| | Objet 2 | Échec, le bucket n'est pas détenu localement | |
| | Objet 3<br>Objet 4<br>Objet 5<br>Objet 6 | Échec<br>Le site en panne peut uniquement lire et mettre à jour les objets détenus localement dans des buckets détenus localement | Success |

Une fois la pulsation rétablie entre les sites, le système signale le site comme étant en ligne et l'accès à ces données se poursuit comme avant la défaillance. L'opération de réintégration :

- Mettra à jour les tableaux de liste des buckets
- Mettra à jour les propriétés d'objets si nécessaire
- Reprendra le traitement de la file d'attente de réplication des sites précédemment en panne

**Remarque :** ECS prend uniquement en charge l'accès lors de la panne temporaire d'un site unique.

Dans une configuration à deux sites, comme illustré à laFigure 13. Les deux sites se considéreront en ligne, signaleront l'autre site comme étant en panne lorsqu'une TSO se produit. Toutes les opérations de création, de liste, de lecture et de mise à jour seront réussies.

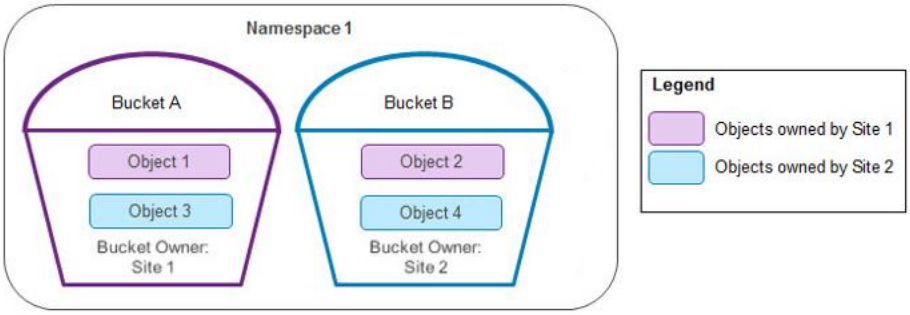


Figure 13 Propriété des buckets et des objets sur deux sites

Au cours de la TSO, tous les objets sont mis à jour sur chaque site. La Tableau 18 affiche les données finales sur le site une fois la pulsation rétablie entre les sites.

Tableau 18 Site gagnant une fois le site rétabli

| Objet | Bucket Name | Bucket Owner | Propriétaire de l'objet | Ensuite, le site gagnant est... |
|---|---|---|---|---|
| Object1 | Bucket A | Site 1 | Site 1 | Site 2 |
| Object2 | Bucket B | Site 2 | Site 1 | Le site dispose de l'horodatage le plus récent |
| Object3 | Bucket A | Site 1 | Site 2 | Le site dispose de l'horodatage le plus récent |
| Object4 | Bucket B | Site 2 | Site 2 | Site 1 |

**Remarque :** dans cet exemple, l'horodatage le plus récent désigne l'heure de la dernière mise à jour de l'objet sur le site.

#### 4.1.2.1 Décodage XOR sur au moins trois sites

Comme nous l'avons vu dans la section Encodage XOR, ECS optimise l'efficacité du stockage des données configurées avec un groupe de réplication contenant au moins trois sites. Les données des copies secondaires des fragments peuvent être remplacées par des données dans un fragment de parité après une opération XOR. Les demandes de données dans un fragment codé seront traitées par le site contenant la copie primaire. Si le site est en panne, la demande est envoyée sur le site avec la copie secondaire de l'objet. Toutefois, puisque cette copie a été codée, le site secondaire doit d'abord récupérer la copie des fragments qui ont été utilisés pour le codage sur les sites principaux en ligne. Ensuite, il effectue une opération XOR pour restituer l'objet demandé et répondre à la demande. Lorsque les fragments sont reconstruits, ils sont également mis en cache afin que le site puisse répondre plus rapidement aux demandes ultérieures.

La Tableau 19 présente un exemple d'une partie d'une table de gestionnaire de fragments sur le Site 4 dans une configuration à quatre sites.

Tableau 19 Table du gestionnaire de fragments du Site 4 après le codage XOR

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 4 | Codé |
| C2 | Site 2 | Site 4 | Codé |
| C3 | Site 3 | Site 4 | Codé |
| C4 | Site 4 | | Parité (C1, C2 et C3) |

Figure 14 illustre les demandes liées à la recréation d'un fragment pour répondre une demande de lecture au cours d'une TSO.

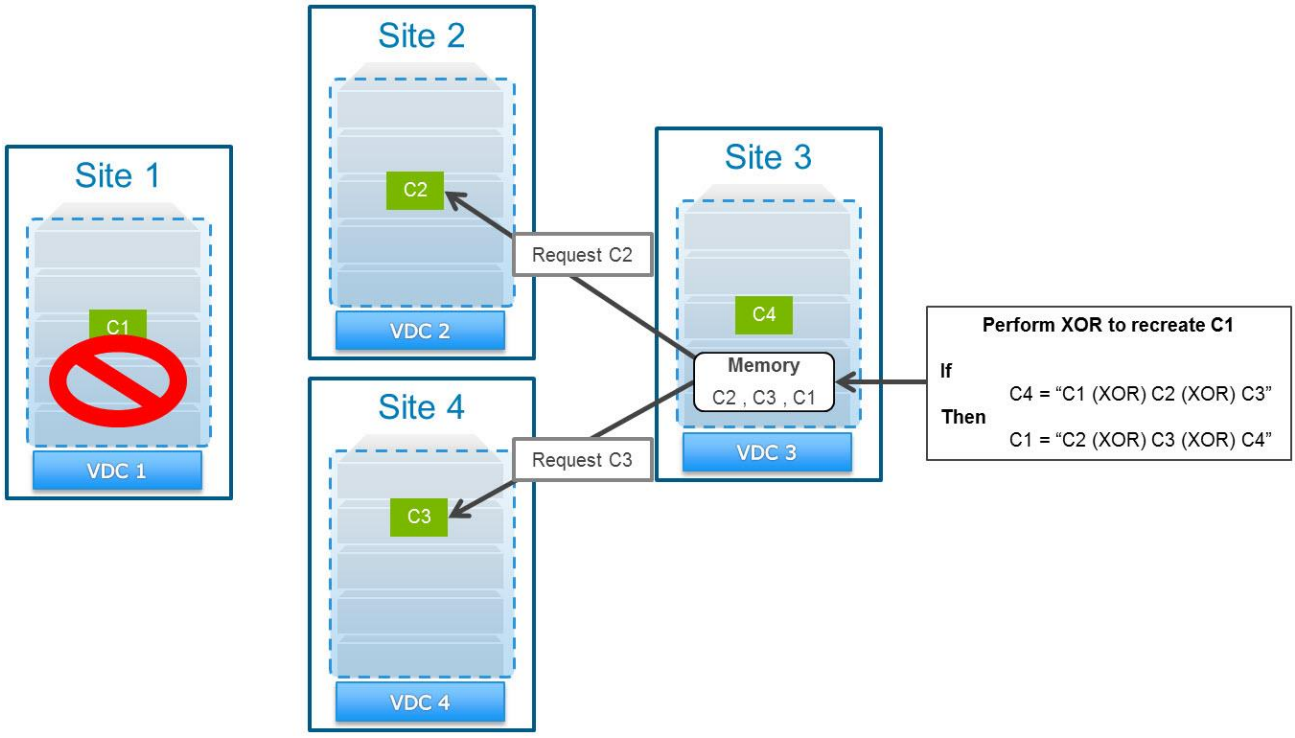


Figure 14 Traitement d'une demande de lecture en reconstruisant un fragment codé par XOR

Dans cet exemple, si une lecture est demandée pour un objet dans le fragment **C1** lorsque le **Site 1** est signalé comme étant en panne, les événements suivants se produisent :

- Puisque le Site 1 est en panne, la demande est envoyée au fragment du site secondaire du C1, Site 4
- Le Site 4 a déjà effectué des opérations XOR sur les fragments **C1, C2** et **C3**, ce qui signifie qu'il a remplacé sa copie locale des données de ces fragments par des données du fragment de parité **C4**.
- Le Site 4 demande une copie du fragment **C2** depuis son site principal (Site 2) et le met en cache localement.
- Le Site 4 demande une copie du fragment **C3** depuis son site principal (Site 3) et le met en cache localement.
- Le Site 4 effectue ensuite une opération XOR entre les fragments mis en cache **C2** et **C3** avec le fragment de parité **C4** pour recréer le fragment **C1** et le stocker localement dans le cache.
- Le Site 4 répond ensuite à la demande de lecture de l'objet dans le fragment **C1**.

**Remarque :** Le temps nécessaire à l'exécution des opérations de reconstruction augmente de manière linéaire en fonction du nombre de sites dans un groupe de réplication.

#### 4.1.2.2 Décodage XOR avec réplication géopassive

Toutes les données d'un bucket configuré avec la réplication géopassive disposeront de deux à quatre sites sources et d'une ou deux cibles de réplication dédiées. Les données des cibles de réplication peuvent être remplacées par des données dans un fragment de parité après une opération XOR. Les demandes de données répliquées géopassivement seront traitées par le site contenant la copie primaire. Si ce site n'est pas accessible au site demandeur, les données doivent être récupérées depuis l'un des sites cibles de réplication.

Avec la réplication géopassive, les sites sources sont toujours les propriétaires d'objets et de buckets. Dans ce cas, si un site cible de réplication est signalé comme temporairement en panne, toutes les opérations d'E/S se poursuivent normalement. La seule exception concerne la réplication qui restera en file d'attente jusqu'à ce que le site cible de réplication rejoigne la fédération.

Si l'un des sites sources échoue, les demandes envoyées au site source en ligne devront restaurer les données qui ne sont pas détenues localement depuis l'un des sites cibles de réplication. Examinons un exemple dans lequel le Site 1 et le Site 2 sont les sites sources et le Site 3 est le site cible de réplication. Dans cet exemple, la copie primaire d'un objet existe dans le fragment C1 qui appartient au Site 1 et le fragment a été répliqué vers la destination cible, le Site 3. Si le Site 1 échoue et qu'une demande arrive sur le Site 2 pour lire cet objet, le Site 2 doit obtenir une copie à partir du Site 3. Si la copie a été codée, le site secondaire doit d'abord récupérer la copie de l'autre fragment qui a été utilisé pour le codage sur le site principal en ligne. Ensuite, il effectue une opération XOR pour restituer l'objet demandé et répondre à la demande. Lorsque les fragments sont reconstruits, ils sont également mis en cache afin que le site puisse répondre plus rapidement aux demandes ultérieures.

La Tableau 20 montre un exemple d'une partie d'une table de gestionnaire de fragments sur la cible de réplication géopassive.

Tableau 20 Table de gestionnaire de fragments de la cible de réplication géopassive après avoir terminé le codage XOR

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Codé |
| C2 | Site 2 | Site 3 | Codé |
| C3 | Site 3 | | Parité (C1 et C2) |

Figure 15 illustre les demandes liées à la recréation d'un fragment pour répondre une demande de lecture au cours d'une TSO.

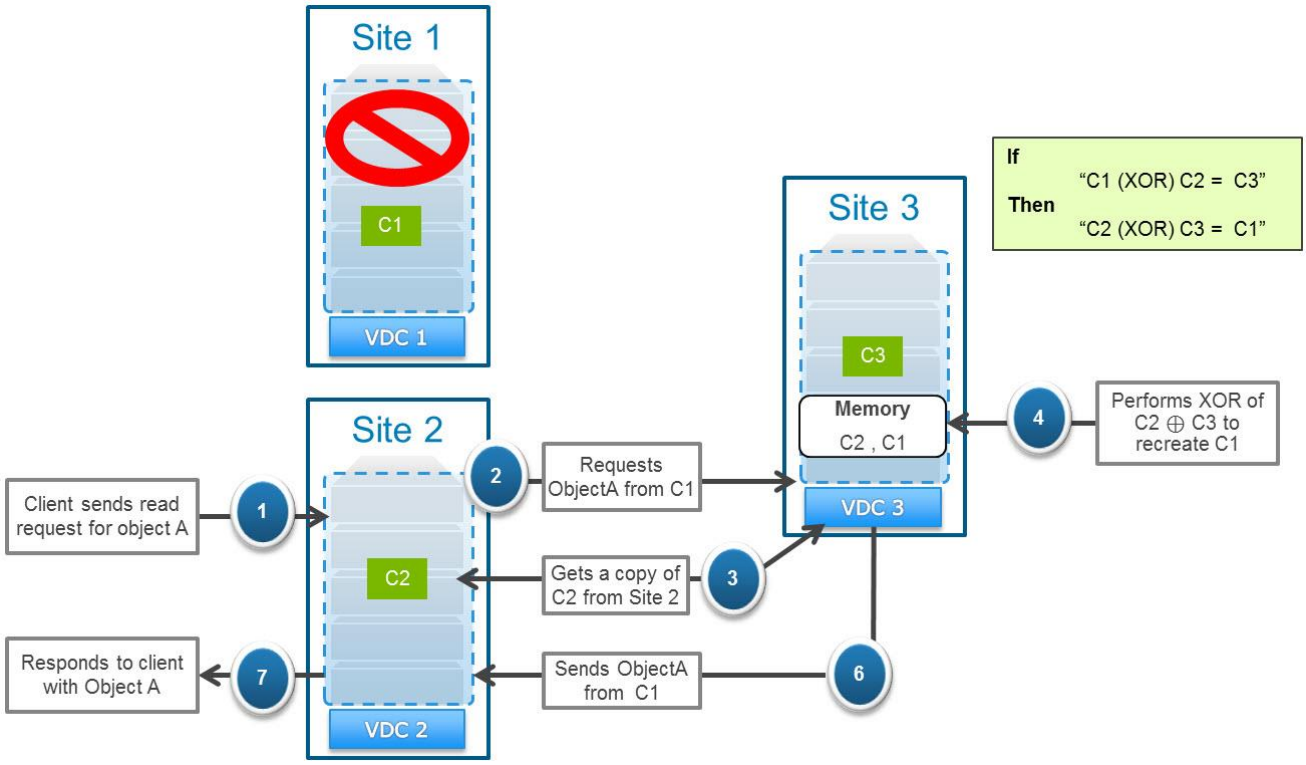


Figure 15 Traitement d'une demande de lecture en reconstruisant un fragment codé par XOR

Dans cet exemple, si une lecture est demandée pour un objet dans le fragment **C1** lorsque le **Site 1** est signalé comme étant temporairement en panne, les événements suivants se produisent :

- Puisque le Site 1 est en panne, la demande est envoyée au fragment du site secondaire du C1, Site 3.
- Le Site 3 a déjà effectué des opérations XOR sur les fragments **C1** et **C2**, ce qui signifie qu'il a remplacé sa copie locale des données de ces fragments par des données du fragment de parité **C3**.
- Le Site 3 demande une copie du fragment **C2** depuis son site principal (Site 2) et le met en cache localement.
- Le Site 3 effectue ensuite une opération XOR entre les fragments mis en cache **C2** et **C3** pour recréer le fragment **C1** et le stocker localement dans le cache.
- Le Site 3 répond ensuite à la demande de lecture de l'objet dans le fragment **C1**.

### 4.1.2.3 Décodage XOR avec la fonction Replicate to all sites activée

Les buckets configurés avec les options **Replicate to all sites** et **Access during outage** peuvent offrir des performances de lecture plus rapides. Ils offrent ces performances de lecture plus rapides à la fois pendant une période où tous les sites sont en ligne, mais également lors d'une panne de site temporaire, car aucune opération de décodage XOR n'est requise et il y a plus de chances que les données soient lues localement.

Les données des buckets avec l'option **Replicate to all sites** activée sont répliquées sur chaque site. La création et la mise à jour d'objets sont gérées de la même manière que si l'option **Replicate to all sites** était désactivée. Toutefois, la lecture et la liste des objets sont gérées légèrement différemment, car certaines données peuvent n'avoir été répliquées que sur certains sites, mais pas sur tous, avant la panne du site principal.

Lors d'une opération de lecture, le nœud qui traite la demande vérifie d'abord la dernière version des métadonnées du propriétaire de l'objet. Si le nœud demandeur :

- **Est le propriétaire de l'objet :**
  - S'il dispose d'une copie locale des données demandées, il l'utilisera pour traiter la demande.
  - Si l'objet a été mis à jour par un autre site qui est tombé en panne avant la réplication des données, il restitue la version qu'il possède localement.
- **N'est pas le propriétaire de l'objet**
  - Si le site propriétaire de l'objet est en ligne et que la réplication des données d'objet :
    - s'est effectuée sur ce site, il traite alors la demande avec sa copie locale des données.
    - ne s'est pas effectuée sur ce site, il demande une copie au propriétaire de l'objet et l'utilise pour traiter la demande.
  - Si le propriétaire de l'objet est en panne
    - Si la réplication de l'objet est effectuée sur ce site, elle utilisera sa copie locale des données. Il peut ne pas s'agir de la dernière version.
    - Si la réplication des données d'objet sur le site demandeur n'a pas abouti, le site demandeur demande une copie du site secondaire répertorié en premier dans la table du gestionnaire de fragments. S'il se trouve être le site secondaire, l'opération de lecture échouera.

Au cours de la création d'une liste d'objets dans un bucket, le nœud nécessite à la fois la création d'une liste d'informations provenant du propriétaire du bucket et les informations en en-tête de chaque objet du bucket. Si le site propriétaire de l'objet ou le propriétaire du bucket est en panne et que l'option **Access During Outage** est également activée, il peut toujours répondre à la demande si tous les sites restants du groupe de réplication sont en ligne. Il répertorie la dernière version du bucket qu'il possède. Elle peut être légèrement obsolète et peut varier d'un site à l'autre.

### 4.1.3 Panne de plusieurs sites

ECS prend uniquement en charge l'accès en cas de panne temporaire d'un site unique au sein d'un groupe de réplication. De plus, un seul site peut être signalé comme étant en panne. Cela signifie que si plusieurs sites au sein d'un groupe de réplication tombent en panne simultanément, certaines opérations échoueront. Le premier site à être considéré comme étant en panne (en raison d'une perte prolongée de pulsation) est signalé comme étant en panne. Tous les sites restants qui accusent également une perte prolongée de pulsation ne seront pas marqués comme étant en panne, et seront donc considérés comme en ligne.

Par exemple, si nous avons cinq sites dans un groupe de réplication et que le Site 1 est identifié comme ayant une perte prolongée de pulsation, il est signalé comme étant en panne. Si le Site 2 est également identifié comme ayant une perte prolongée de pulsation, il reste répertorié comme étant en ligne.
Les événements suivants se produisent :

- Si l'option **Access During Outage** est activée et que le propriétaire du bucket est le Site 2, les lectures/créations/mises à jour envoyées à d'autres sites échoueront, quel que soit le propriétaire de l'objet. C'est ce qu'il se produit parce qu'il vérifie d'abord auprès du propriétaire du bucket pour déterminer le propriétaire de l'objet. Sauf si le propriétaire du bucket est signalé comme étant en panne, le demandeur envoie la demande au Site 2, qui échoue.
- Les demandes de lecture et de mise à jour envoyées au Site 2 n'aboutissent que s'il s'agit du propriétaire de l'objet (et du propriétaire du bucket si l'option **Access During Outage** est activée).

- Les demandes de lecture et de mise à jour envoyées à d'autres sites que le Site 2 n'aboutissent que s'il s'agit du propriétaire de l'objet (et du propriétaire du bucket si l'option **Access During Outage** est activée) n'est pas le Site 2.
- La création de l'objet échoue si le propriétaire du bucket est Site 1 ou Site 2. Cela s'explique par la nécessité pour la création d'un objet de la mise à jour de la liste de buckets avec le nouveau nom d'objet. Étant donné qu'il ne peut pas réussir à effectuer cette opération sur tous les sites signalés en ligne, l'opération de création échouera.
- Les demandes de création d'une liste d'objets dans un bucket n'aboutissent que si le site demandeur peut accéder au propriétaire du bucket et à tous les objets.
  - Si la demande est envoyée au Site 2, elle n'aboutira que s'il est propriétaire du bucket et de tous les objets du bucket.
  - Si la demande est envoyée à un autre site, elle ne réussira que si ni le bucket ni aucun objet du bucket ne sont détenus par le Site 2.

## 4.2 Panne de site permanente (PSO)

Si un sinistre se produit sur un site et que l'administrateur détermine que le site est irrécupérable, il peut lancer un basculement de site permanent (supprimer le VDC de la fédération). Lorsqu'un basculement de site permanent est lancé, tous les fragments du site en échec sont récupérés sur les sites restants pour rétablir la durabilité des données.

Le processus de récupération implique que les sites restants analysent leur table de gestionnaire de fragments local à la recherche de références aux sites qui incluent le site en échec. Tout ce qu'il trouve avec un type de fragment de :

- **Codé**
  a. Pour les fragments dont le type est codé et dont le site principal est en ligne, il recréera les données localement à l'aide des données du site principal. Une fois terminé, ce fragment est marqué comme étant une copie.
  b. Ensuite, il recrée le fragment codé dont le site principal est le site en panne en effectuant une opération XOR des fragments de copies précédemment recréés avec le fragment de parité. Ce site devient désormais le site principal des fragments et devient un site local.
  c. Ces fragments sont ensuite ajoutés à la file d'attente de réplication pour être répliqués vers d'autres sites répertoriés dans le groupe de réplication.

- Le fragment **Copy** et un site principal répertorié en tant que site en panne deviennent le nouveau site principal. Il ajoute ensuite le fragment à sa file d'attente de réplication pour le répliquer sur un nouveau site secondaire.
- Le fragment **Local** et son site secondaire sont répertoriés comme le site en panne. Une tâche est insérée pour répliquer le fragment vers un nouveau site secondaire.

Une fois le basculement de site permanent démarré, l'accès aux données détenues par le site en panne ne sera pas disponible avant la fin du processus de basculement permanent du site. La réplication des données est distincte des opérations de basculement et, par conséquent, elle n'a pas besoin d'être terminée pour que l'accès aux données détenues par le site en échec soit restauré.

Si l'on examine un exemple à trois sites, le Site 1 tombe en panne et les Tableau 21 et Tableau 22 sont les tables de gestionnaire de fragments des deux sites restants.

Tableau 21 Table du gestionnaire de fragments du Site 2

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Copy |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 1 | Site 3 | Remote |
| C4 | Site 2 | Site 1 | Local |

Le Site 2 effectue les opérations suivantes :

- Ajoute le fragment **C1** à la file d'attente de réplication pour être répliqué. Le Site 2 deviendra le nouveau site principal et le site avec le nouveau fragment deviendra le nouveau site secondaire.
- Ajoute le fragment **C4** à la file d'attente de réplication pour répliquer et mettre à jour le site secondaire dans la table.

Tableau 22 Table du gestionnaire de fragments du Site 3

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 2 | Remote |
| C2 | Site 2 | Site 3 | Codé |
| C3 | Site 1 | Site 3 | Codé |
| C4 | Site 2 | Site 1 | Remote |
| C5 | Site 3 | | Parité (C2 et C3) |

Le Site 3 effectue les opérations suivantes :

1. Recrée les données du fragment **C2** localement à l'aide des données du site principal (**Site 2**). Change le type de fragment en Copie.
2. Reconstruit le fragment **C3** à l'aide des données de **C2** et des données de parité **C5** en effectuant une opération XOR C2 ⊕ C5. Le Site 3 deviendra le nouveau site principal.
3. Supprime le fragment **C5**.
4. Ajoute le fragment **C3** à la file d'attente de réplication pour le répliquer. Le site détenant le nouveau fragment deviendra le nouveau site secondaire.

Une fois que le basculement de site permanent a complété le gestionnaire de fragments, les tables des deux sites restants sont comme dans les Tableau 23 et Tableau 24.

Tableau 23 Table du gestionnaire de fragments du Site 2 après la fin d'une PSO.

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 2 | Site 3 | Local |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 3 | Site 2 | Copy |
| C4 | Site 2 | Site 3 | Local |

Tableau 24 Table du gestionnaire de fragments du Site 3 après la fin d'une PSO

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 2 | Site 3 | Copy |
| C2 | Site 2 | Site 3 | Copy |
| C3 | Site 3 | Site 2 | Local |
| C4 | Site 2 | Site 3 | Copy |

### 4.2.1 PSO avec réplication géopassive

Les pannes de site permanentes sont gérées légèrement différemment pour les données répliquées géopassivement. Les données répliquées géopassivement ne rétabliront pas la durabilité des données au cours d'une PSO. Au lieu de cela, elle est rétablie après l'ajout d'un troisième site au groupe de réplication. Les opérations PSO diffèrent selon que le site qui a échoué de manière permanente est l'un des sites sources ou le site cible de réplication.

Le processus de récupération implique toujours que les sites restants analysent leur table de gestionnaire de fragments local à la recherche de références aux sites qui incluent le site en échec. Tout ce qu'il trouve avec un type de fragment de :

- **Codé** (existe sur le site cible de réplication)
  - Pour les fragments dont le type est codé et dont le site principal est en ligne, il recréera les données localement à l'aide des données du site principal. Une fois terminé, ce fragment est marqué comme étant une copie.
  - Ensuite, il recrée le fragment codé dont le site principal est le site source en panne en effectuant une opération XOR des fragments de copies précédemment recréés avec le fragment de parité. Ce site devient désormais le site principal des fragments et devient un site local. Aucun site secondaire ne sera créé avant l'ajout d'un troisième site au groupe de réplication.
- Le fragment **Copy** et un site principal répertorié en tant que site en panne deviennent le nouveau site primaire et deviennent un site local. Aucun site secondaire ne sera créé avant l'ajout d'un troisième site au groupe de réplication.
- Le fragment **Local** et son site secondaire (la cible de réplication) sont le site en panne. Aucun nouveau site secondaire ne sera créé avant l'ajout d'un troisième site au groupe de réplication.

Après une PSO, un troisième site peut être ajouté pour rétablir la durabilité des données et se protéger contre les pannes à l'échelle du site. Après l'ajout d'un troisième site au groupe de réplication géopassive, les deux sites précédents analysent leur table de gestionnaire de fragments local à la recherche de fragments sans fragment secondaire répertorié. Ensuite :

- Les fragments locaux sur un site source sans fragment secondaire répertorié lancent la réplication d'un fragment secondaire vers le nouveau site cible de réplication. La table du gestionnaire de fragments sera mise à jour pour inclure le nouveau site du fragment secondaire.
- Les fragments locaux sur le site cible de réplication lancent une réplication du fragment vers un nouveau site source. Une fois la réplication terminée, le type de site cible de réplication passe de Local à Copy, et le type de site source passe de Copy à Local. Les opérations XOR se poursuivent normalement vers le site cible.

Une fois la PSO démarrée sur un site source, l'accès aux données détenues par le site en panne ne sera pas disponible avant la fin du processus de basculement permanent du site. Une fois la PSO terminée, l'accès aux données est restauré. Tant qu'un troisième site n'est pas ajouté au groupe de réplication, toutes les nouvelles écritures sur le site source en ligne sont répliquées sur la cible de réplication, mais les opérations XOR ne se produisent pas. Étant donné que tous les nouveaux sites sources sont identiques, et que XOR s'exécute uniquement sur des fragments de deux sites sources différents, XOR ne peut pas s'exécuter.

Une fois la PSO terminée, un troisième site peut être ajouté au groupe de réplication pour rétablir la durabilité des données et se protéger contre les pannes à l'échelle du site. De plus, la cible de réplication peut reprendre l'exécution des opérations XOR sur deux fragments provenant de sites sources différents marqués comme copie.

Examinons quelques exemples dans lesquels le Site 1 et le Site 2 sont les sites sources et le Site 3 est le site cible. Les Tableau 25 et Tableau 26 sont les tables de gestionnaire de fragments des deux sites sources, et la Tableau 27 est la table de gestionnaire de fragments du site cible de réplication.

Tableau 25 Site 1, table du gestionnaire de fragments du site source

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Local |
| C2 | Site 2 | Site 3 | Remote |
| C3 | Site 1 | Site 3 | Local |

Tableau 26 Site 2, table du gestionnaire de fragments du site source

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Remote |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 1 | Site 3 | Remote |

Tableau 27 Site 3, table du gestionnaire de fragments de la cible de réplication

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Codé |
| C2 | Site 2 | Site 3 | Codé |
| C3 | Site 1 | Site 3 | Copy |
| C4 | Site 3 | | Parité (C1 et C2) |

**Exemple 1 :** Si le Site 3 est supprimé en raison d'une PSO, les sites secondaires deviennent tous vides, mais les sites et types principaux restent. Tant qu'une nouvelle cible de réplication n'est pas ajoutée, toute nouvelle écriture aura un site principal répertorié, mais aucun site secondaire.

**Exemple 2 :** Si Site1 est supprimé en raison d'une PSO, les événements suivants se produisent :

- Le Site 3 devient le nouveau site principal avec un fragment local « **C3** », et aucun site ne sera répertorié comme site secondaire.
- Le Site 3 recrée les données du fragment **C2** à l'aide des données du site principal (Site 2) et modifiera son type de fragment en Copy.
- Le Site 3 reconstruit le fragment **C1** à l'aide des données de **C2** et des données de parité **C4** en effectuant une opération XOR C2 ⊕ C4. Le Site 3 devient le nouveau site principal ; aucun site secondaire ne sera répertorié.
- Le Site 3 supprime le fragment **C4**.

Une fois que le basculement de site permanent du Site 1 terminé, les tables de gestionnaire de fragments des deux sites restants sont comme dans les Tableau 28 et Tableau 29.

Tableau 28 Site 2, table du gestionnaire de fragments du site source après la fin d'une PSO.

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 3 | | Remote |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 3 | | Remote |

Tableau 29 Site 3, table du gestionnaire de fragments du site cible de réplication après la fin d'une PSO

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 3 | | Local |
| C2 | Site 2 | Site 3 | Copy |
| C3 | Site 3 | | Local |

Tant qu'un nouveau site source n'est pas ajouté, toutes les nouvelles écritures disposeront d'un site principal de Site 2 avec un type de site local et d'un site secondaire de Site 3 avec un type de copie.

Après l'ajout d'un nouveau site source au groupe de réplication, la durabilité des données afin de se protéger contre les pannes à l'échelle du site sera rétablie en ajoutant un site secondaire et en y répliquant les données. Les opérations XOR reprennent également sur la cible de réplication. Les nouvelles tables de gestionnaire de fragments sont indiquées dans les Tableau 30 à Tableau 32.

- Les fragments C1 et C3 seront répliqués vers le nouveau site source, Site 1. Une fois la réplication terminée, le site principal est répertorié en tant que Site 1 et le site secondaire en tant que Site 3.
- Le Site 3 effectue l'encodage XOR sur les fragments C1 et C2, ce qui entraîne la création d'un nouveau fragment C4, de parité. Les fragments C1 et C2 deviennent des segments encodés.

Tableau 30 Nouvelle table de gestionnaire de fragments du Site 1 après le rétablissement de la durabilité des données

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Local |
| C2 | Site 2 | Site 3 | Remote |
| C3 | Site 1 | Site 3 | Local |

Tableau 31 Table du gestionnaire de fragments du Site 2 après l'ajout du nouveau Site 1 et du rétablissement de la durabilité des données

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Remote |
| C2 | Site 2 | Site 3 | Local |
| C3 | Site 1 | Site 3 | Remote |

Tableau 32 Table du gestionnaire de fragments de la cible de réplication du Site 3 après l'ajout du nouveau Site 1 et du rétablissement de la durabilité des données

| ID de fragment | Site principal | Site secondaire | Type |
|---|---|---|---|
| C1 | Site 1 | Site 3 | Codé |
| C2 | Site 2 | Site 3 | Codé |
| C3 | Site 1 | Site 3 | Copy |
| C4 | Site 3 | | Parité (C1 et C2) |

### 4.2.2 Capacité de récupération en cas de pannes sur plusieurs sites

ECS prend uniquement en charge la récupération d'une panne de site à la fois. ECS peut effectuer une récupération après plusieurs pannes de site si les opérations de PSO et de récupération des données se terminent entre les pannes du site. Si le deuxième site échoue avant la fin de la récupération :

- Tous les autres sites du système doivent être en ligne pour que les opérations de récupération après une panne de site permanente s'exécutent. En cas de pannes simultanées de plusieurs sites, tous les sites, sauf un, doivent se remettre de la TSO avant qu'une PSO puisse être exécutée sur un site.
- Si une deuxième panne du site se produit après la fin de la PSO, mais avant la fin de la récupération des données, certaines données peuvent être perdues.

Prenons comme exemple un scénario à quatre sites dans lequel nous avons réussi à nous remettre de la perte de tous les sites sauf un (en supposant qu'il y ait suffisamment d'espace pour stocker toutes les données sur le site restant) :

- Le Site 4 tombe en panne
  - L'administrateur lance une opération de PSO pour supprimer le Site 4
  - Les données sont récupérées sur les sites restants pour rétablir la durabilité des données

  Il nous reste maintenant une fédération à trois sites contenant Site 1, Site 2 et Site 3.

- Le deuxième site tombe en panne, Site 2 :

  Parfois, un autre site peut échouer, par exemple Site 2, après la PSO et la récupération des données.

  - L'administrateur lance une opération de PSO pour supprimer le Site 2
  - Les données sont récupérées sur les sites restants pour rétablir la durabilité des données

Il nous reste maintenant une fédération à deux sites contenant Site 1 et Site 3.

- Un troisième site tombe en panne, Site 1 :

  Parfois, un autre site peut échouer, par exemple Site 1, après la PSO et la récupération des données.

  - L'administrateur lance une opération de PSO pour supprimer le Site 1

  Il nous reste désormais une fédération à site unique contenant le Site 3.

Il s'agit d'un exemple avec plusieurs pannes de site. Ce n'est pas un scénario courant. Les pannes de site permanentes sont généralement causées par des sinistres, tels que les tremblements de terre ou les incendies, et il n'est pas courant qu'elles se produisent sur plusieurs sites les uns à la suite des autres. En général, après une panne permanente d'un seul site, un nouveau site est ajouté avant qu'un site ne tombe en panne.

# 5 Conclusion

L'architecture ECS a été conçue dès le départ pour offrir à la fois disponibilité du système et durabilité des données. ECS offre à l'administrateur de la granularité dans la manière d'équilibrer les exigences de disponibilité et le coût TCO. Des fonctionnalités telles que la détection automatique des pannes et l'autoréparation réduisent les charges applicatives pour les équipes informatiques aux moments les plus critiques, lorsqu'il y a un événement non planifié, tel qu'une panne de site.

ECS protège les données d'un site ou d'un VDC contre les pannes de disques à l'aide d'une combinaison de mise en miroir triple et de codage d'effacement. ECS offre deux niveaux de protection par codage d'effacement, la valeur par défaut pour les cas d'utilisation classiques et le stockage à froid, plus efficace pour les objets rarement consultés. Il répartit également les données entre les domaines de panne afin de fournir une protection contre la plupart des scénarios de pannes.

ECS garantit l'intégrité des données en calculant et en écrivant des sommes de contrôle dans le cadre d'une opération d'écriture et en validant ces sommes de contrôle lors d'une opération de lecture. La validation de la somme de contrôle est également effectuée proactivement dans une tâche en arrière-plan.

ECS est conçu pour continuer à assurer la disponibilité du système. Tout cela est possible avec l'architecture distribuée qui permet de traiter les demandes client par n'importe quel nœud d'un site ou d'un VDC.

La conception d'ECS étend la disponibilité du système et la protection de la durabilité des données en ajoutant une protection facultative contre une panne complète à l'échelle du site. Pour ce faire, il fédère les sites et permet à l'administrateur de configurer diverses options de règles de groupe de réplication. Ces options peuvent être définies au niveau du bucket et déterminer où répliquer les données, comment stocker les données sur le ou les sites distants, ainsi que l'option Access During Outage.

De plus, ECS offre aux clients une option « Access During Outage », ce qui permet de lire, répertorier et éventuellement d'écrire et de mettre à jour des opérations envoyées sur un site en ligne lorsque le bucket et/ou l'objet sont signalés comme étant en panne.

Si un administrateur détermine qu'un site est irrécupérable, il peut déclencher une panne de site permanente. Cela supprime le VDC ou le site du groupe de réplication et recrée les données en fonction des besoins pour rétablir la durabilité des données.

En conclusion, ECS offre une solution de stockage dans le Cloud de qualité professionnelle avec une résilience intégrée à laquelle vous pouvez faire confiance.

# A Support technique et ressources

Dell.com/support propose des services et un support éprouvés répondant aux besoins des clients.

Des documents et vidéos techniques sur le stockage offrent aux clients l'expertise nécessaire pour tirer pleinement parti des plates-formes de stockage Dell EMC.

## A.1 Ressources associées

- Livre blanc : présentation et architecture de la solution ECS
- Communauté ECS
- ECS Test Drive
- Documentation sur les produits ECS sur le site de support ou sur le site de la communauté
- SolVe Desktop (Procedure Generator)