



Collecte des données et contrôle des performances des workloads

OnCommand Unified Manager 9.5

NetApp
December 20, 2023

Sommaire

Collecte des données et contrôle des performances des workloads	1
Types de charges de travail surveillés par Unified Manager	1
Valeurs de mesure des performances des charges de travail	2
Parmi les performances attendues	5
Comment la plage attendue est utilisée dans l'analyse des performances	6
Comment Unified Manager utilise une latence de charge de travail pour identifier les problèmes de performance	7
Comment les opérations d'un cluster peuvent affecter la latence des charges de travail	8
Contrôle des performances des configurations MetroCluster	8
Analyse et notification des événements de performance	11
Comment Unified Manager détermine l'impact sur les performances d'un événement	13
Les composants du cluster et les conflits	14
Rôles des charges de travail impliquées dans un événement de performance	15

Collecte des données et contrôle des performances des workloads

Unified Manager collecte et analyse les activités des charges de travail toutes les 5 minutes afin d'identifier les événements de performances et détecte les changements de configuration toutes les 15 minutes. Il conserve un maximum de 30 jours de données d'historique des performances et des événements de 5 minutes. Ces données permettent de prévoir la plage attendue pour toutes les charges de travail surveillées.



Ce chapitre décrit le fonctionnement des seuils dynamiques et leur utilisation pour surveiller les performances des charges de travail. Ce chapitre ne s'applique pas aux statistiques ou aux événements provoqués par les violations de seuil de performance définies par l'utilisateur ou par le système.

Unified Manager doit collecter au moins 3 jours d'activité sur la charge de travail avant de pouvoir lancer l'analyse. Il doit avant que la plage prévue pour le temps de réponse et les opérations d'E/S ne soit affichée sur la page Détails des performances/volumes et sur la page Détails de l'événement. Pendant la collecte de cette activité, la plage attendue n'affiche pas toutes les modifications effectuées à partir de l'activité de la charge de travail. Après avoir collecté 3 jours d'activité, Unified Manager ajuste la plage attendue, toutes les 24 heures à 12 h 00, pour prendre en compte les modifications de l'activité des charges de travail et établir un seuil de performances plus précis.

Au cours des 4 premiers jours suivant la surveillance d'un volume par Unified Manager, si plus de 24 heures ont été écoulées depuis la dernière collecte de données, les graphiques de la page Performance/Volume Details n'affichent pas la plage prévue pour ce volume. Les événements détectés avant la dernière collection sont toujours disponibles.



L'heure d'été change l'heure système, ce qui modifie la plage attendue de statistiques de performances pour les charges de travail surveillées. Unified Manager commence immédiatement à corriger la plage attendue, qui prend environ 15 jours. Pendant cette période, vous pouvez continuer à utiliser Unified Manager, mais, puisque Unified Manager utilise la plage attendue pour détecter des événements, il se peut que certains événements ne soient pas précis. Les événements détectés avant le changement de temps ne sont pas affectés. La modification manuelle de l'heure sur un cluster ou sur un serveur Unified Manager à une heure antérieure affecte également les résultats de l'analyse d'événements.

Types de charges de travail surveillés par Unified Manager

Unified Manager permet de surveiller les performances de deux types de charges de travail : définies par l'utilisateur et défini par le système.

- **charges de travail définies par l'utilisateur**

Débit d'E/S des applications vers le cluster. Ce sont des processus impliqués dans les requêtes de lecture et d'écriture. Un volume FlexVol ou FlexGroup est une charge de travail définie par l'utilisateur.



Unified Manager surveille uniquement l'activité des workloads sur le cluster. Il ne surveille pas les applications, les clients ou les chemins d'accès entre les applications et le cluster.

Si un ou plusieurs des éléments suivants sont vrais pour une charge de travail, il ne peut pas être surveillé par Unified Manager :

- Il s'agit d'une copie de protection des données (DP) en mode lecture seule. (Notez que lors de l'utilisation de ONTAP 8.3 ou version ultérieure, les volumes DP sont surveillés pour le trafic généré par les utilisateurs.)
- C'est un Infinite Volume.
- Il s'agit d'un clone de données hors ligne.
- Il s'agit d'un volume en miroir dans une configuration MetroCluster.

• **charges de travail définies par le système**

Les processus internes impliqués dans l'efficacité du stockage, la réplication des données et l'état du système, notamment :

- Efficacité du stockage, comme la déduplication
- État de santé du disque, qui inclut la reconstruction RAID, le nettoyage du disque, etc
- Réplication des données, notamment les copies SnapMirror
- Activités de gestion
- État de santé du système de fichiers, qui inclut les diverses activités WAFL
- Les scanners de système de fichiers, tels que la numérisation WAFL
- Allègement de la charge des copies, comme les opérations d'efficacité du stockage déchargées depuis les hôtes VMware
- État du système, comme les déplacements de volume, la compression des données, etc
- Volumes non surveillés

Les données de performance des charges de travail définies par le système s'affichent dans l'interface graphique uniquement lorsque le composant de cluster utilisé par ces charges de travail conflits. Par exemple, vous ne pouvez pas rechercher le nom d'une charge de travail définie par le système pour afficher les données de performance dans l'interface graphique. Si plusieurs charges de travail du même type sont affichées, une lettre est ajoutée au nom du workload. La lettre est destinée au personnel de soutien.

Valeurs de mesure des performances des charges de travail


Unified Manager mesure les performances des workloads sur un cluster en fonction des valeurs statistiques historiques et attendues, qui constituent la plage de valeurs attendue pour les workloads. Il compare les valeurs statistiques réelles de la charge de travail à la plage prévue pour déterminer si les performances des charges de travail sont trop élevées ou trop faibles. Un workload qui n'exécute pas comme prévu entraîne l'envoi d'un rapport d'événements de performance à vous notifier.

Dans l'illustration suivante, la valeur réelle, en rouge, représente les statistiques de performance réelles dans la période. La valeur réelle a dépassé le seuil de performances, qui est les limites supérieures de la plage attendue. Le pic est la valeur réelle la plus élevée dans la période. L'écart mesure le changement entre les valeurs attendues et les valeurs réelles, tandis que l'écart de crête indique le changement le plus important entre les valeurs attendues et les valeurs réelles.



Le tableau suivant répertorie les valeurs de mesure des performances des workloads.

Mesure	Description
Activité	<p>Pourcentage de la limite de qualité de service utilisée par les workloads dans le groupe de règles.</p> <p>Si Unified Manager détecte une modification au groupe de règles, par exemple l'ajout ou la suppression d'un volume ou la modification de la limite de QoS, les valeurs réelles et attendues peuvent dépasser 100 % de la limite définie. Si une valeur dépasse 100 % de la limite définie, elle s'affiche sous la forme de > 100 %. Si une valeur est inférieure à 1 % de la limite définie, elle s'affiche sous la forme <1 %.</p>
Réel	La valeur des performances mesurée à un moment spécifique pour une charge de travail donnée.

Mesure	Description
Déviation	<p>Changement entre les valeurs attendues et les valeurs réelles. Il s'agit du rapport entre la valeur réelle moins la valeur attendue et la valeur supérieure de la plage attendue moins la valeur attendue.</p> <div style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;">  <p>Une valeur de déviation négative indique que la performance de la charge de travail est inférieure à la valeur attendue, tandis qu'une valeur de déviation positive indique que la performance de la charge de travail est supérieure à la valeur attendue. Si les valeurs attendues et la valeur réelle sont très faibles, dans les centièmes ou les millièmes de pourcentage par exemple, la déviation affiche N/A.</p> </div>
Attendu	<p>Les valeurs attendues sont basées sur l'analyse des données historiques de performances pour une charge de travail donnée. Unified Manager analyse ces valeurs statistiques afin de déterminer la plage de valeurs attendue.</p>
Plage prévue	<p>La plage de valeurs attendue est une prévision, ou prévision, de ce que les valeurs de performance supérieure et inférieure devraient être à un moment donné. Pour la latence des workloads, les valeurs supérieures constituent le seuil de performance. Lorsque la valeur réelle franchit le seuil de performances, Unified Manager déclenche une alerte d'événement de performance.</p>
Pic	<p>Valeur maximale mesurée sur une période de temps.</p>
Déviation de crête	<p>Valeur de déviation maximale mesurée sur une période de temps.</p>
Profondeur de la file d'attente	<p>Nombre de demandes d'E/S en attente du composant d'interconnexion.</p>

Mesure	Description
Du stockage	Pour les composants de traitement de réseau, de traitement de données et d'agrégat, le pourcentage de temps d'activité requis pour mener à bien les opérations de la charge de travail sur une période donnée. Par exemple, le pourcentage de temps alloué aux composants de traitement réseau ou de traitement de données pour traiter une demande d'E/S ou à un agrégat pour répondre à une demande de lecture ou d'écriture.
Débit d'écriture	Débit en écriture, en mégaoctets par seconde (Mbit/s), des charges de travail sur un cluster local vers le cluster partenaire dans une configuration MetroCluster.

Parmi les performances attendues

La plage de valeurs attendue est une prévision, ou prévision, de ce que les valeurs de performance supérieure et inférieure devraient être à un moment donné. Pour la latence des workloads, les valeurs supérieures constituent le seuil de performance. Lorsque la valeur réelle franchit le seuil de performances, Unified Manager déclenche une alerte d'événement de performance.

Par exemple, pendant les heures de bureau habituelles, entre 9:00 à 5 h 00, la plupart des employés peuvent vérifier leur courriel entre 9 h 00 et 10:30 L'augmentation de la demande sur les serveurs de messagerie entraîne une augmentation de l'activité de la charge de travail sur le stockage interne au cours de cette période. Les employés risquent de remarquer le ralentissement des temps de réponse de la part de leurs clients de messagerie.

Pendant l'heure du déjeuner, entre 12:00 et 13:00 et à la fin de la journée de travail après 5 h 00, la plupart des employés sont susceptibles de s'éloigner de leurs ordinateurs. La demande sur les serveurs de messagerie diminue généralement, tout en diminuant la demande sur le stockage interne. Il peut également y avoir des opérations planifiées pour les charges de travail, telles que les sauvegardes de stockage ou l'analyse antivirus, commençant après 5 h 00 et augmenter l'activité sur le stockage interne.

Sur plusieurs jours, l'augmentation et la diminution de l'activité de la charge de travail déterminent l'étendue d'activité attendue, avec des limites supérieure et inférieure pour une charge de travail. Lorsque l'activité de workload réelle d'un objet se trouve en dehors des limites supérieure ou inférieure et reste en dehors des limites pendant un certain temps, il peut indiquer que l'objet est sur-utilisé ou sous-utilisé.

La manière dont la plage attendue est formée

Unified Manager doit collecter au moins 3 jours d'activité sur la charge de travail avant de pouvoir lancer l'analyse et avant que la plage prévue pour le temps de réponse d'E/S et les opérations ne puissent être affichées dans l'interface graphique. La collecte de données minimale requise ne tient pas compte de toutes les modifications apportées à l'activité de la charge de travail. Une fois l'activité des 3 premiers jours terminée, Unified Manager ajuste la plage attendue, toutes les 24 heures à 12 h 00. Il reflète les modifications apportées à l'activité des charges de travail et établit un seuil de performances plus précis.



L'heure d'été change l'heure système, ce qui modifie la plage attendue de statistiques de performances pour les charges de travail surveillées. Unified Manager commence immédiatement à corriger la plage attendue, qui prend environ 15 jours. Pendant cette période, vous pouvez continuer à utiliser Unified Manager, mais, puisque Unified Manager utilise la plage attendue pour détecter des événements, il se peut que certains événements ne soient pas précis. Les événements détectés avant le changement de temps ne sont pas affectés. La modification manuelle de l'heure sur un cluster ou sur un serveur Unified Manager à une heure antérieure affecte également les résultats de l'analyse d'événements.

Comment la plage attendue est utilisée dans l'analyse des performances

Unified Manager utilise la plage attendue pour représenter la latence d'E/S (temps de réponse) et l'activité d'opérations d'entrée/sortie par seconde (opérations) types pour vos charges de travail surveillées. Elle vous avertit lorsque la latence réelle d'une charge de travail se situe au-dessus des limites supérieures de la plage attendue, ce qui déclenche un événement de performance. Vous pouvez ainsi analyser le problème de performance et prendre des mesures correctives pour le résoudre.

La plage attendue définit la ligne de base des performances pour la charge de travail. Il apprend des mesures de performance passées dont il a besoin pour prévoir les niveaux de performance et d'activité attendus pour la charge de travail. La limite supérieure de la plage attendue établit le seuil de performance. Unified Manager utilise le modèle de base pour déterminer quand la ou les opérations réelles se situent au-dessus ou en dessous d'un seuil ou en dehors des limites de leur plage prévue. La comparaison entre les valeurs réelles et attendues crée un profil de performances pour la charge de travail.

Lorsque la latence réelle d'une charge de travail dépasse le seuil de performance défini en raison d'un conflit au niveau d'un composant du cluster, la latence est élevée et la charge de travail fonctionne plus lentement que prévu. Les performances des autres charges de travail qui partagent les mêmes composants du cluster peuvent également être plus lentes que prévu.

Unified Manager analyse l'événement seuil à atteindre et détermine si l'activité est un événement de performances. Si l'activité élevée des charges de travail reste cohérente pendant une longue période, telles que plusieurs heures, Unified Manager considère que l'activité est normale et ajuste de manière dynamique la plage attendue pour définir le nouveau seuil de performance.

Certaines charges de travail peuvent avoir une activité faible et cohérente, dans laquelle les opérations ou la latence n'est pas fortement modifiée. Pour réduire le nombre d'alertes d'événements, lors de l'analyse des événements de performances, Unified Manager déclenche un événement uniquement pour les volumes à faible activité dont les opérations et la latence sont supérieures à celles que nous attendions.



Dans cet exemple, la latence d'un volume est une plage attendue, en gris, de 0 milliseconde par opération (ms/op) à sa plus basse et de 5 ms/op à sa plus élevée. Si la latence réelle, en bleu, augmente soudainement à 10 ms/opération, en raison d'un pic intermittent du trafic réseau ou de conflits sur un composant du cluster, il

est alors supérieur à la plage prévue et a dépassé le seuil de performance.

Lorsque le trafic réseau a diminué ou que le composant du cluster n'est plus en conflit, la latence renvoie à la plage attendue. Si la latence reste supérieure ou égale à 10 ms/opération sur une longue période, vous pouvez être contraint d'effectuer une action corrective pour résoudre le problème.

Comment Unified Manager utilise une latence de charge de travail pour identifier les problèmes de performance

La latence (temps de réponse) correspond au temps nécessaire pour qu'un volume d'un cluster réponde aux demandes d'E/S des applications client. Unified Manager utilise la latence pour détecter les événements de performance et vous alerter.

Une latence élevée signifie que les demandes provenant des applications vers un volume d'un cluster prennent plus de temps que d'habitude. La cause de la latence élevée peut se trouver sur le cluster lui-même, en raison d'un conflit sur un ou plusieurs composants du cluster. Une latence élevée peut également être provoquée par des problèmes en dehors du cluster, tels que des goulets d'étranglement du réseau, des problèmes avec le client qui héberge les applications ou des problèmes avec ces mêmes applications.



Unified Manager surveille uniquement l'activité des workloads sur le cluster. Il ne surveille pas les applications, les clients ou les chemins d'accès entre les applications et le cluster.

Les opérations sur le cluster, comme effectuer des sauvegardes ou exécuter une déduplication, qui augmentent les besoins des composants de cluster partagés par d'autres charges de travail peuvent également contribuer à la latence élevée. Si la latence réelle dépasse le seuil de performances de la plage attendue, Unified Manager analyse l'événement afin de déterminer s'il s'agit d'un événement de performances auquel vous devrez peut-être résoudre. La latence est mesurée en millisecondes par opération (ms/op).

La page Performance/Volume Details vous permet d'afficher une analyse des statistiques de latence afin de voir comment l'activité de chaque processus, comme les demandes de lecture et d'écriture, se compare aux statistiques de latence globales. La comparaison vous permet de déterminer quelles opérations ont l'activité la plus élevée ou si des opérations spécifiques ont une activité anormale qui affecte la latence d'un volume. Lors de l'analyse des événements de performances, vous pouvez utiliser les statistiques de latence pour déterminer si un événement a été provoqué par un problème sur le cluster. Vous pouvez également identifier les activités spécifiques à la charge de travail ou les composants de cluster impliqués dans l'événement.



Cet exemple montre le graphique latence sur la page Performance/Volume Details. L'activité du temps de réponse réel (latence) est une ligne bleue et la plage attendue est grise.

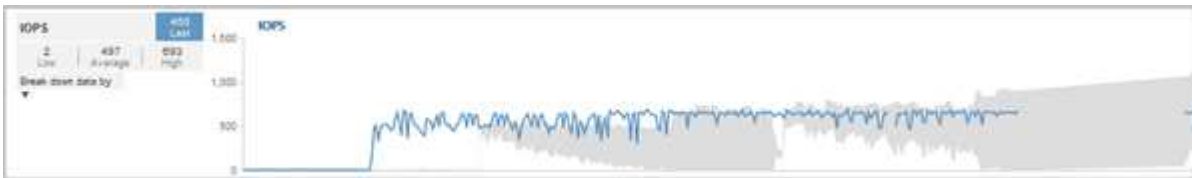


Il peut y avoir des lacunes dans la ligne bleue si Unified Manager n'a pas pu collecter des données. Cela peut se produire du fait que le cluster ou le volume était inaccessible, Unified Manager a été désactivé pendant cette période ou que la collecte a pris plus de 5 minutes.

Comment les opérations d'un cluster peuvent affecter la latence des charges de travail

Les opérations (IOPS) représentent l'activité de tous les workloads définis par le système et l'utilisateur sur un cluster. Les statistiques Op E/S par sec vous aident à déterminer si les processus du cluster, tels que réaliser des sauvegardes ou exécuter la déduplication, ont un impact sur la latence (temps de réponse) d'une charge de travail ou s'ils ont pu être responsables d'un événement de performances.

Lors de l'analyse des événements de performances, vous pouvez utiliser les statistiques relatives aux IOPS pour déterminer si un événement de performances a été provoqué par un problème sur le cluster. Vous pouvez identifier les activités spécifiques à chaque charge de travail qui peuvent être les principales sources d'événements de performances. Les IOPS sont mesurées en opérations par seconde (OPS/s).



Cet exemple montre le graphique IOPS sur la page Performance/Volume Details. Les statistiques d'opérations réelles sont une ligne bleue et la plage de statistiques d'opérations prévue est grise.



Dans certains cas où un cluster est surchargé, Unified Manager peut afficher le message `Data collection is taking too long on Cluster cluster_name`. Cela signifie que les statistiques à analyser sont insuffisantes pour Unified Manager. Vous devez réduire les ressources utilisées par le cluster afin de collecter les statistiques.

Contrôle des performances des configurations MetroCluster

Unified Manager vous permet de contrôler le débit d'écriture entre les clusters d'une configuration MetroCluster afin d'identifier les workloads dont le débit d'écriture est élevé. Si ces charges de travail hautes performances provoquent d'autres volumes du cluster local des temps de réponse d'E/S élevés, Unified Manager déclenche des événements de performance qui vous avertissent.

Lorsqu'un cluster local d'une configuration MetroCluster met en miroir ses données vers son cluster partenaire, les données sont écrites sur la mémoire NVRAM, puis transférées sur les liens ISL vers les agrégats distants. Unified Manager analyse la mémoire NVRAM pour identifier les charges de travail dont le débit d'écriture élevé utilise la mémoire NVRAM, ce qui engendre des conflits.

Les charges de travail dont la déviation dans le temps de réponse a dépassé le seuil de performance sont appelées *victimes* et les charges de travail dont l'écart dans le débit d'écriture vers la NVRAM est plus élevé que d'habitude, entraînant la contention, sont appelées *bullies*. Seules les demandes d'écriture sont mises en miroir vers le cluster partenaire, Unified Manager n'analyse pas le débit de lecture.

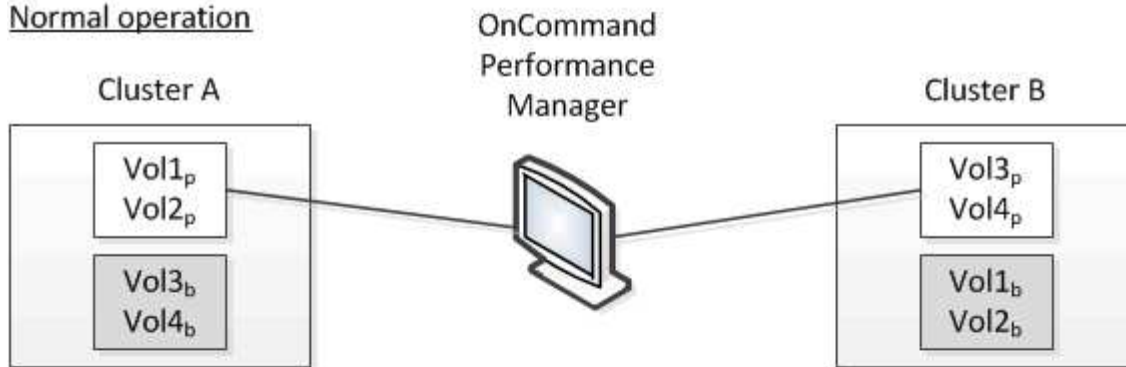
Unified Manager traite les clusters dans une configuration MetroCluster comme des clusters individuels. Il ne fait aucune distinction entre les clusters qui sont des partenaires ou établit un lien entre le débit d'écriture de chaque cluster.

Comportement des volumes lors du basculement et du rétablissement

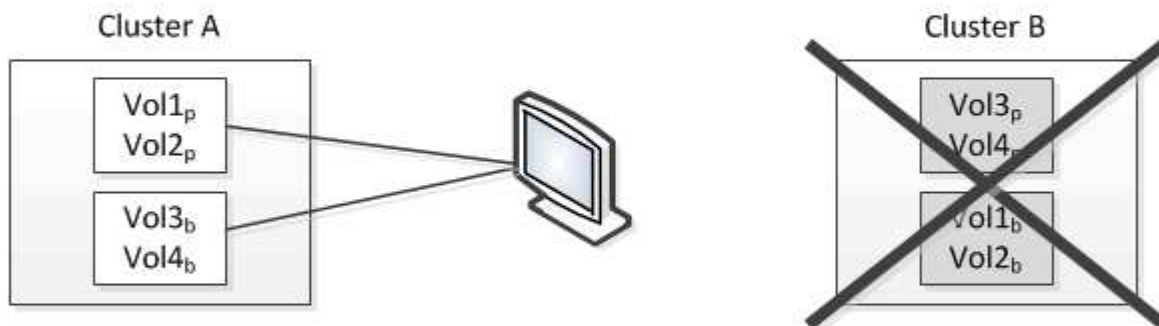
Les événements qui déclenchent un basculement ou un rétablissement entraînent le déplacement des volumes actifs d'un cluster vers l'autre cluster du groupe de reprise sur incident. Les volumes du cluster qui étaient actifs et devaient transmettre des données aux clients sont arrêtés, et les volumes de l'autre cluster sont activés et commencent à transmettre les données. Unified Manager surveille uniquement les volumes actifs et en cours d'exécution.

Comme les volumes sont déplacés d'un cluster à l'autre, il est recommandé de contrôler les deux clusters. Une seule instance de Unified Manager peut contrôler les deux clusters dans une configuration MetroCluster, mais parfois la distance entre les deux sites nécessite l'utilisation de deux instances Unified Manager pour surveiller les deux clusters. La figure suivante présente une seule instance de Unified Manager :

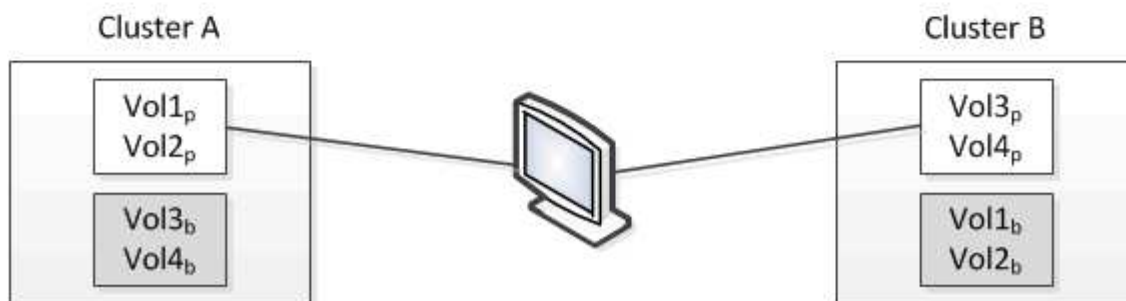
Normal operation



Cluster B fails --- switchover to Cluster A



Cluster B is repaired --- switchback to Cluster B



 = active and monitored by OPM  = inactive and not monitored by OPM

Les volumes portant la référence p indiquent les volumes primaires, et les volumes dont l'nom est b sont des volumes de sauvegarde en miroir créés par SnapMirror.

En fonctionnement normal :

- Le cluster A a deux volumes actifs : Vol1p et Vol2p.
- Le cluster B a deux volumes actifs : Vol3p et Vol4p.
- Cluster A comporte deux volumes inactifs : Vol3b et Vol4b.
- Le cluster B a deux volumes inactifs : Vol1b et Vol2b.

Les informations relatives à chacun des volumes actifs (statistiques, événements, etc.) sont collectées par Unified Manager. Les statistiques Vol1p et Vol2p sont collectées par le Cluster A et les statistiques Vol3p et Vol4p sont recueillies par le Cluster B.

Après une défaillance majeure, entraîne le basculement des volumes actifs du Cluster B vers le Cluster A :

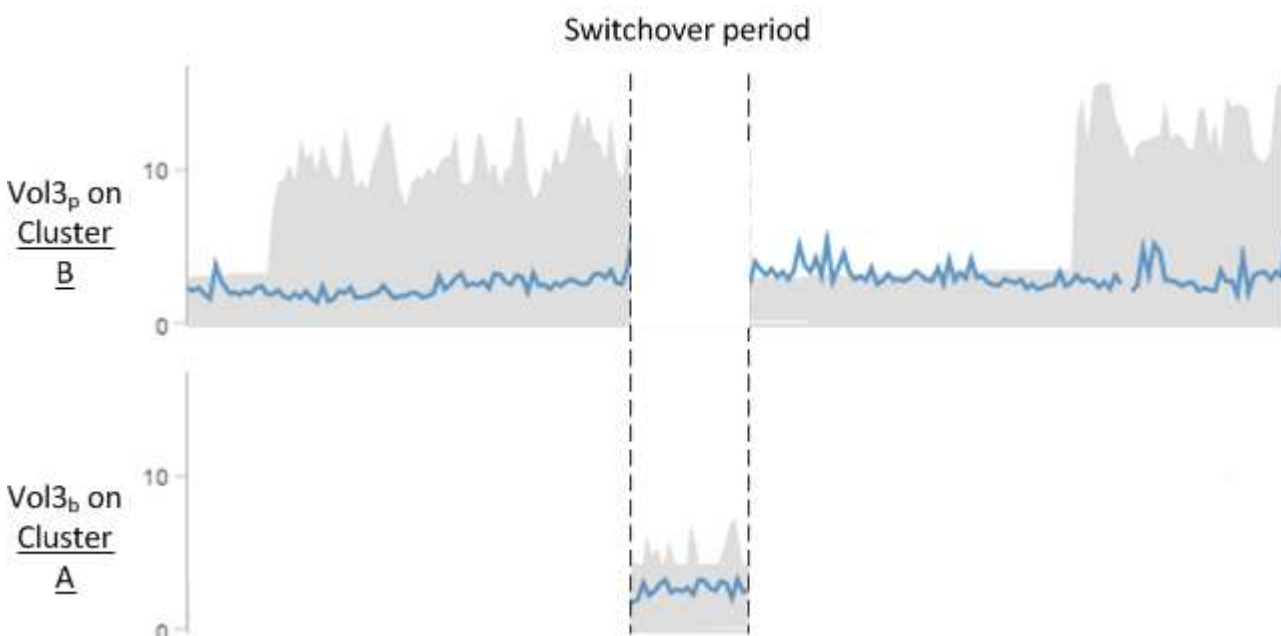
- Cluster A contient quatre volumes actifs : Vol1p, Vol2p, Vol3b et Vol4b.
- Le cluster B a quatre volumes inactifs : Vol3p, Vol4p, Vol1b et Vol2b.

Comme pendant le fonctionnement normal, les informations relatives à chacun des volumes actifs sont collectées par Unified Manager. Mais dans ce cas, les statistiques Vol1p et Vol2p sont recueillies par le Cluster A, et les statistiques Vol3b et Vol4b sont également recueillies par le Cluster A.

Notez que Vol3p et Vol3b ne sont pas les mêmes volumes, car ils se trouvent sur des clusters différents. Les informations contenues dans Unified Manager pour Vol3p ne sont pas les mêmes que Vol3b :

- Lors du basculement vers le Cluster A, les statistiques et les événements Vol3p ne sont pas visibles.
- Lors du premier basculement, Vol3b ressemble à un nouveau volume sans informations historiques.

Lorsque le Cluster B est réparé et qu'un rétablissement est effectué, Vol3p est de nouveau actif sur le Cluster B, avec les statistiques historiques et un intervalle de statistiques correspondant à la période de basculement. Vol3b n'est pas visible depuis le Cluster A tant qu'un autre basculement se produit :





- Ainsi, les volumes MetroCluster inactifs, Vol3b sur le Cluster A après rétablissement, sont identifiés par le message « ce volume a été supprimé ». Le volume n'est pas supprimé, mais n'est actuellement pas surveillé par Unified Manager, car il ne s'agit pas du volume actif.
- Lorsqu'un seul Unified Manager contrôle les deux clusters dans une configuration MetroCluster, la recherche de volume renvoie les informations correspondant au volume actif à ce moment-là. Par exemple, une recherche de « Vol3 » renverrait des statistiques et des événements pour Vol3b sur le Cluster A si un basculement s'est produit et Vol3 est devenu actif sur le Cluster A.

Analyse et notification des événements de performance

Les événements de performance vous signalent les problèmes de performances d'E/S au niveau d'une charge de travail de volume causée par des conflits dans un composant de cluster. Unified Manager analyse l'événement pour identifier toutes les charges de travail impliquées, le composant dans les conflits et si l'événement reste un problème à résoudre.

Unified Manager surveille la latence (temps de réponse) et les IOPS (opérations) des volumes d'un cluster. Lorsque d'autres charges de travail surfont un composant de cluster, par exemple, les conflits sont possibles et le composant ne peut pas fonctionner à un niveau optimal pour répondre aux demandes de charge de travail. Les performances des autres charges de travail qui utilisent le même composant peuvent être affectées, ce qui entraîne une augmentation des latences. Si la latence franchit le seuil de performance, Unified Manager déclenche un événement de performance et envoie une alerte e-mail vous avertie.

Analyse des événements

Unified Manager effectue les analyses suivantes, en s'appuyant sur les statistiques de performance des 15 derniers jours, pour identifier les workloads victime, les workloads dominants et le composant de cluster impliqué dans un événement :

- Identifie les charges de travail victimes dont la latence a dépassé le seuil de performance, qui est la limite supérieure de la plage attendue :
 - Pour les volumes des agrégats HDD ou Flash Pool (hybride), les événements sont déclenchés uniquement lorsque la latence dépasse 5 millisecondes (ms) et que les IOPS représentent plus de 10 opérations par seconde (OPS/s).
 - Pour les volumes situés sur des agrégats 100 % SSD ou des agrégats FabricPool (composites), les événements sont déclenchés uniquement lorsque la latence est supérieure à 1 ms et que les IOPS sont plus de 100 OPS/s.
- Identifie le composant de cluster dans les conflits.



Si la latence des charges de travail victimes au niveau de l'interconnexion de cluster est supérieure à 1 ms, Unified Manager le traite comme important et déclenche un événement pour l'interconnexion de cluster.

- Identifie les charges de travail dominantes qui font l'objet d'une surutilisation du composant de cluster et qui l'entraînent des conflits.
- Classe les charges de travail impliquées, en fonction de leur déviation de l'utilisation ou de l'activité d'un composant du cluster, afin de déterminer les principaux changements d'utilisation du composant du cluster

et les victimes les plus affectées.

Un événement peut se produire brièvement et se corriger après le composant qu'il utilise n'est plus en conflit. Un événement continu est un événement qui se produit de nouveau pour le même composant de cluster au cours d'un intervalle de cinq minutes et qui reste à l'état actif. Pour les événements continus, Unified Manager déclenche une alerte après avoir détecté le même événement à deux intervalles d'analyse consécutifs. Les événements qui ne sont pas résolus, qui ont un état nouveau, peuvent afficher différents messages de description en tant que charges de travail impliquées dans le changement d'événement.

Lorsqu'un événement est résolu, il reste disponible dans Unified Manager dans le cadre de l'enregistrement des anciens problèmes de performances d'un volume. Chaque événement possède un ID unique qui identifie le type d'événement et les volumes, le cluster et les composants de cluster impliqués.



Un seul volume peut être impliqué dans plusieurs événements simultanément.

État de l'événement

Les événements peuvent être dans l'un des États suivants :

- **Actif**

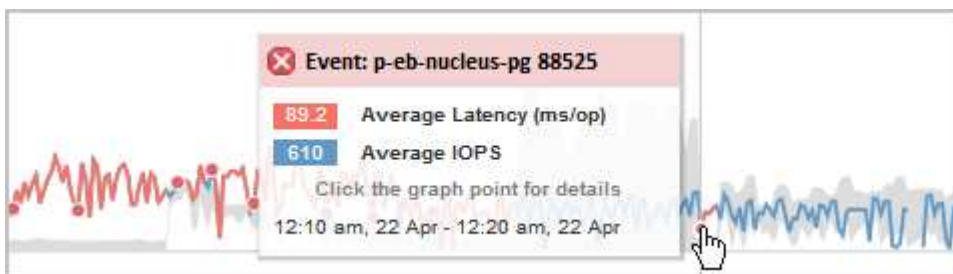
Indique que l'événement de performance est actuellement actif (nouveau ou reconnu). Le problème à l'origine de l'incident n'a pas été corrigé lui-même ou n'a pas été résolu. Le compteur de performances de l'objet de stockage reste au-dessus du seuil de performance.

- **Obsolète**

Indique que l'incident n'est plus actif. Le problème à l'origine de l'incident s'est corrigé ou a été résolu. Le compteur de performance de l'objet de stockage n'est plus au-dessus du seuil de performance.

Notification d'événement

Les alertes d'événement s'affichent sur la page tableaux de bord/Présentation, tableaux de bord/performance, page Détails performances/Volume et sont envoyées à des adresses e-mail spécifiées. Vous pouvez afficher des informations d'analyse détaillées sur un événement et obtenir des suggestions de résolution de cet événement sur la page Détails de l'événement.



Dans cet exemple, un événement est indiqué par un point rouge (●) Sur le graphique latence de la page Détails performances/Volume. Lorsque vous déplacez le curseur de la souris sur le point rouge, une fenêtre contextuelle contenant plus de détails sur l'événement et les options d'analyse s'affiche.

Interaction d'événement

Sur la page Performance/Volume Details, vous pouvez interagir avec les événements de l'une des manières suivantes :

- Le déplacement du pointeur sur un point rouge affiche un message indiquant l’ID de l’événement, la latence, le nombre d’opérations par seconde, ainsi que la date et l’heure de détection de l’événement.

S’il y a plusieurs événements pour la même période, le message affiche le nombre d’événements, ainsi que la latence moyenne et les opérations par seconde pour le volume.

- Lorsque vous cliquez sur un seul événement, une boîte de dialogue affiche des informations plus détaillées sur l’événement, notamment les composants de cluster impliqués, similaires à la section Résumé de la page Détails de l’événement.

Le composant en conflit est entouré et mis en évidence en rouge. Vous pouvez cliquer sur l’ID d’événement ou sur **Afficher l’analyse complète** pour afficher l’analyse complète sur la page Détails de l’événement. S’il existe plusieurs événements pour la même période, la boîte de dialogue affiche des détails sur les trois événements les plus récents. Vous pouvez cliquer sur un ID d’événement pour afficher l’analyse d’événement sur la page Détails de l’événement. S’il y a plus de trois événements pour la même période, le fait de cliquer sur le point rouge n’affiche pas la boîte de dialogue.

Comment Unified Manager détermine l’impact sur les performances d’un événement

Unified Manager utilise l’écart d’activité, d’utilisation, de débit d’écriture, de l’utilisation d’un composant du cluster ou de latence d’E/S (temps de réponse) pour une charge de travail afin de déterminer le niveau d’impact sur les performances d’une charge de travail. Ces informations déterminent le rôle de chaque charge de travail dans l’événement et leur classement sur la page Détails de l’événement.

Unified Manager compare les dernières valeurs analysées pour une charge de travail à la plage de valeurs attendue. La différence entre les valeurs analysées pour la dernière fois et la plage de valeurs attendue identifie les workloads pour lesquels les performances ont le plus été affectées par l’événement.

Supposons par exemple qu’un cluster contienne deux charges de travail : la charge De travail A et la charge de travail B. La plage prévue pour la charge de travail A est de 5-10 millisecondes par opération (ms/op) et sa latence réelle est généralement d’environ 7 ms/op. La plage attendue pour la charge de travail B est de 10-20 ms/op et sa latence réelle est généralement d’environ 15 ms/op. La latence prévue pour les deux charges de travail est très bonne. En raison des conflits sur le cluster, la latence des deux charges de travail augmente à 40 ms/activité, ce qui atteint le seuil de performances, qui correspond aux limites supérieures de la plage attendue et aux événements déclencheurs. L’écart de latence, entre les valeurs attendues et les valeurs supérieures au seuil de performances, pour la charge de travail A est d’environ 33 ms/op, et l’écart pour la charge de travail B est d’environ 25 ms/op. La latence des deux charges de travail atteint 40 ms/activité, mais la charge de travail A avait l’impact le plus important sur les performances, car elle avait l’écart de latence le plus élevé à 33 ms/opération.

Sur la page Détails de l’événement, dans la section diagnostic système, vous pouvez trier les charges de travail par variation de l’activité, de l’utilisation ou du débit d’un composant de cluster. Vous pouvez également trier les charges de travail par latence. Lorsque vous sélectionnez une option de tri, Unified Manager analyse l’écart en termes d’activité, d’utilisation, de débit ou de latence depuis que l’événement a été détecté à partir des valeurs attendues pour déterminer l’ordre de tri de la charge de travail. Pour la latence, les points rouges (●) indiquent un seuil de performances franchissement par une charge de travail victime et l’impact qui en découle sur la latence. Chaque point rouge indique un niveau d’écart plus élevé de latence, ce qui vous aide à identifier les workloads victimes dont la latence a le plus été affectée par un événement.

Les composants du cluster et les conflits

Vous pouvez identifier les problèmes de performance du cluster lorsqu'un composant du cluster entre en conflit. Les performances des charges de travail de volumes qui utilisent le composant ralentissent et leur temps de réponse (latence) augmente pour les requêtes client, ce qui déclenche un événement dans Unified Manager.

Un composant en conflit ne peut pas se faire à un niveau optimal. Ses performances ont diminué, et la performance des autres composants et charges de travail du cluster, appelés *victimes*, peut avoir augmenté la latence. Pour mettre un composant à l'extérieur des conflits, vous devez réduire sa charge de travail ou augmenter sa capacité à gérer davantage de travail, de sorte que les performances puissent revenir à des niveaux normaux. Unified Manager collecte et analyse les performances des charges de travail toutes les cinq minutes. En effet, il ne détecte que lorsqu'un composant du cluster est constamment sur-utilisé. Les pics transitoires de surutilisation qui durent pendant une courte durée dans l'intervalle de cinq minutes ne sont pas détectés.

Par exemple, un agrégat de stockage peut être soumis à des conflits car une ou plusieurs charges de travail y sont en concurrence pour que leurs demandes d'E/S soient traitées. Des charges de travail peuvent être affectées sur l'agrégat, ce qui entraîne une baisse des performances. Pour réduire la quantité d'activité sur l'agrégat, plusieurs étapes sont possibles : déplacer une ou plusieurs charges de travail vers un agrégat moins occupé, par exemple, réduire la demande de charge de travail globale sur l'agrégat actuel. Pour un groupe de règles de qualité de service, vous pouvez ajuster la limite de débit ou déplacer les workloads vers un autre groupe de règles, de sorte que les charges de travail ne soient plus restreintes.

Unified Manager contrôle les composants de cluster suivants pour vous alerter en cas de conflit :

- **Réseau**

Représente le temps d'attente des demandes d'E/S par les protocoles iSCSI ou Fibre Channel (FC) sur le cluster. Le temps d'attente correspond au temps passé à attendre la fin des transactions iSCSI Ready to Transfer (R2T) ou FCP Transfer Ready (XFER_RDY) pour que le cluster puisse répondre à une demande d'E/S. Si le composant réseau est en conflit, cela signifie qu'un temps d'attente élevé au niveau de la couche de protocole de bloc a un impact sur la latence d'une ou de plusieurs charges de travail.

- **Traitement réseau**

Composant logiciel dans le cluster impliqué dans le traitement des E/S entre la couche de protocole et le cluster. Le traitement du réseau de traitement des nœuds a peut-être changé depuis la détection de l'événement. Si le composant de traitement de réseau est en conflit, son utilisation élevée au niveau du nœud de traitement réseau a un impact sur la latence d'une ou de plusieurs charges de travail.

- **Politique de qualité de service**

Représente le groupe de règles de qualité de service (QoS) de stockage dont la charge de travail est membre. Si le composant de groupe de règles conflits, cela signifie que toutes les charges de travail du groupe de règles sont restreintes par la limite de débit définie, qui a un impact sur la latence d'une ou plusieurs de ces charges de travail.

- * Interconnexion de cluster*

La représente les câbles et adaptateurs avec lesquels les nœuds en cluster sont physiquement connectés. Si le composant d'interconnexion de cluster est en conflit, cela signifie un temps d'attente élevé pour les demandes d'E/S au niveau de l'interconnexion de cluster se répercute sur la latence d'une ou de plusieurs charges de travail.

- **Traitement de données**

Composant logiciel dans le cluster impliqué dans le traitement des E/S entre le cluster et l'agrégat de stockage qui contient la charge de travail. Le traitement des données de traitement du nœud peut avoir changé depuis la détection de l'événement. Si le composant de traitement des données conflit, une utilisation élevée au niveau du nœud de traitement des données affecte la latence d'un ou de plusieurs workloads.

- **Ressources MetroCluster**

La représente les ressources MetroCluster, y compris la NVRAM et les liens ISL, utilisés pour mettre en miroir les données entre les clusters dans une configuration MetroCluster. Si le composant MetroCluster rencontre des conflits, il s'agit d'un débit d'écriture élevé avec les charges de travail sur le cluster local ou d'un problème d'état de santé de la liaison ayant un impact sur la latence d'une ou de plusieurs charges de travail sur le cluster local. Si le cluster ne se trouve pas dans une configuration MetroCluster, cette icône n'est pas affichée.

- **Agrégat ou agrégat SSD**

Agrégat de stockage sur lequel les charges de travail s'exécutent. Si le composant de l'agrégat est en conflit, une utilisation élevée de l'agrégat a un impact sur la latence d'une ou de plusieurs charges de travail. Un agrégat se compose de tous les disques durs ou d'un mélange de disques durs et de disques SSD (un agrégat Flash Pool). Un « agrégat SD » se compose de tous les SSD (un agrégat 100 % Flash), ou d'une combinaison de SSD et d'un niveau cloud (un agrégat FabricPool).

- * Latence cloud*

Représente le composant logiciel du cluster impliqué dans le traitement des E/S entre le cluster et le niveau cloud sur lequel les données utilisateur sont stockées. Si le composant de latence dans le cloud conflits, une grande quantité de lectures sur les volumes hébergés sur le Tier cloud ont une incidence sur la latence d'un ou de plusieurs workloads.

- **SnapMirror de synchronisation**

Représente le composant logiciel du cluster impliqué dans la réplication des données utilisateur depuis le volume primaire vers le volume secondaire dans une relation SnapMirror synchrone. Si le composant SnapMirror synchrone entre en conflit, l'activité des opérations SnapMirror synchrone a un impact sur la latence d'un ou de plusieurs workloads.

Rôles des charges de travail impliquées dans un événement de performance

Unified Manager utilise des rôles pour identifier la participation d'une charge de travail en cas de performance. Les rôles sont les victimes, les taureaux et les requins. Une charge de travail définie par l'utilisateur peut être une victime, un tyran et un requin en même temps.

Rôle	Description
Victime	Charge de travail définie par l'utilisateur dont les performances ont diminué en raison des autres charges de travail, appelées « bullies », qui sont sur-utilisées lors de l'utilisation d'un composant du cluster. Seules les charges de travail définies par l'utilisateur sont identifiées comme victimes. Unified Manager identifie les charges de travail victimes en fonction de leur écart de latence, où la latence réelle, pendant un événement, a été considérablement améliorée par rapport à la plage de latence prévue.
Intimider	Une charge de travail définie par l'utilisateur ou définie par le système dont l'utilisation excessive d'un composant de cluster a entraîné une diminution des performances d'autres charges de travail, appelées « victimes ». Unified Manager identifie les workloads dominants en fonction de leur déviation par l'utilisation d'un composant de cluster, où l'utilisation réelle, au cours d'un événement, a considérablement augmenté à partir de sa plage d'utilisation prévue.
Requin	Charge de travail définie par l'utilisateur, avec l'utilisation la plus élevée d'un composant de cluster, et non pas toutes les charges de travail impliquées dans un événement. Unified Manager identifie les charges de travail Shark en fonction de leur utilisation d'un composant de cluster pendant un événement.

Les charges de travail d'un cluster peuvent partager la plupart des composants du cluster, tels que les agrégats de stockage et la CPU pour le traitement du réseau et des données. Lorsqu'une charge de travail, par exemple un volume, augmente l'utilisation d'un composant de cluster au point que le composant ne peut pas répondre efficacement aux exigences de la charge de travail, le composant engendre des conflits. La charge de travail sur-utilisation d'un composant de cluster est un phénomène tyran. Les autres charges de travail qui partagent ces composants, et dont la performance est impactée par le tyran, sont les victimes. L'activité provenant des charges de travail définies par le système, telles que la déduplication ou les copies Snapshot, peut également créer des « brimades ».

Lorsqu'Unified Manager détecte un événement, il identifie tous les workloads et composants de cluster impliqués, notamment les workloads dominants qui ont causé l'événement, le composant de cluster en conflit et les workloads victimes dont les performances ont diminué en raison de l'augmentation de l'activité des workloads dominants.



Si Unified Manager ne peut pas identifier les charges de travail dominantes, cette alerte s'applique uniquement aux charges de travail victimes et au composant de cluster concerné.

Unified Manager est capable d'identifier les charges de travail victimes de charges de travail dominantes. Il peut également y avoir une identification lorsque ces mêmes charges de travail deviennent des charges de travail dominantes. Un workload peut être un tyran à lui-même. Par exemple, une charge de travail élevée au ralenti par une limite de groupe de règles entraîne la restriction de toutes les charges de travail du groupe de règles, y compris de celles-ci. Une charge de travail dominante ou victime dans un événement de performance

continu peut changer son rôle ou ne plus y participer. Sur la page Détails performances/Volume, dans le tableau liste d'événements, lorsque le volume sélectionné change son rôle de participant, la date et l'heure du changement de rôle s'affichent.

Informations sur le copyright

Copyright © 2023 NetApp, Inc. Tous droits réservés. Imprimé aux États-Unis. Aucune partie de ce document protégé par copyright ne peut être reproduite sous quelque forme que ce soit ou selon quelque méthode que ce soit (graphique, électronique ou mécanique, notamment par photocopie, enregistrement ou stockage dans un système de récupération électronique) sans l'autorisation écrite préalable du détenteur du droit de copyright.

Les logiciels dérivés des éléments NetApp protégés par copyright sont soumis à la licence et à l'avis de non-responsabilité suivants :

CE LOGICIEL EST FOURNI PAR NETAPP « EN L'ÉTAT » ET SANS GARANTIES EXPRESSES OU TACITES, Y COMPRIS LES GARANTIES TACITES DE QUALITÉ MARCHANDE ET D'ADÉQUATION À UN USAGE PARTICULIER, QUI SONT EXCLUES PAR LES PRÉSENTES. EN AUCUN CAS NETAPP NE SERA TENU POUR RESPONSABLE DE DOMMAGES DIRECTS, INDIRECTS, ACCESSOIRES, PARTICULIERS OU EXEMPLAIRES (Y COMPRIS L'ACHAT DE BIENS ET DE SERVICES DE SUBSTITUTION, LA PERTE DE JOUISSANCE, DE DONNÉES OU DE PROFITS, OU L'INTERRUPTION D'ACTIVITÉ), QUELLES QU'EN SOIENT LA CAUSE ET LA DOCTRINE DE RESPONSABILITÉ, QU'IL S'AGISSE DE RESPONSABILITÉ CONTRACTUELLE, STRICTE OU DÉLICTEUELLE (Y COMPRIS LA NÉGLIGENCE OU AUTRE) DÉCOULANT DE L'UTILISATION DE CE LOGICIEL, MÊME SI LA SOCIÉTÉ A ÉTÉ INFORMÉE DE LA POSSIBILITÉ DE TELS DOMMAGES.

NetApp se réserve le droit de modifier les produits décrits dans le présent document à tout moment et sans préavis. NetApp décline toute responsabilité découlant de l'utilisation des produits décrits dans le présent document, sauf accord explicite écrit de NetApp. L'utilisation ou l'achat de ce produit ne concède pas de licence dans le cadre de droits de brevet, de droits de marque commerciale ou de tout autre droit de propriété intellectuelle de NetApp.

Le produit décrit dans ce manuel peut être protégé par un ou plusieurs brevets américains, étrangers ou par une demande en attente.

LÉGENDE DE RESTRICTION DES DROITS : L'utilisation, la duplication ou la divulgation par le gouvernement sont sujettes aux restrictions énoncées dans le sous-paragraphe (b)(3) de la clause Rights in Technical Data-Noncommercial Items du DFARS 252.227-7013 (février 2014) et du FAR 52.227-19 (décembre 2007).

Les données contenues dans les présentes se rapportent à un produit et/ou service commercial (tel que défini par la clause FAR 2.101). Il s'agit de données propriétaires de NetApp, Inc. Toutes les données techniques et tous les logiciels fournis par NetApp en vertu du présent Accord sont à caractère commercial et ont été exclusivement développés à l'aide de fonds privés. Le gouvernement des États-Unis dispose d'une licence limitée irrévocable, non exclusive, non cessible, non transférable et mondiale. Cette licence lui permet d'utiliser uniquement les données relatives au contrat du gouvernement des États-Unis d'après lequel les données lui ont été fournies ou celles qui sont nécessaires à son exécution. Sauf dispositions contraires énoncées dans les présentes, l'utilisation, la divulgation, la reproduction, la modification, l'exécution, l'affichage des données sont interdits sans avoir obtenu le consentement écrit préalable de NetApp, Inc. Les droits de licences du Département de la Défense du gouvernement des États-Unis se limitent aux droits identifiés par la clause 252.227-7015(b) du DFARS (février 2014).

Informations sur les marques commerciales

NETAPP, le logo NETAPP et les marques citées sur le site <http://www.netapp.com/TM> sont des marques déposées ou des marques commerciales de NetApp, Inc. Les autres noms de marques et de produits sont des marques commerciales de leurs propriétaires respectifs.