Collecte des données et contrôle des performances des workloads

Active IQ Unified Manager 9.7

NetApp
April 17, 2024

Sommaire

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 jusqu'à 30 jours de données d'historique des performances et des événements de 5 minutes. Ces données permettent d'établir les prévisions de latence pour toutes les charges de travail surveillées.

Unified Manager doit collecter au moins 3 jours d'activité de la charge de travail avant de pouvoir lancer son analyse. Pour ce faire, il est possible d'afficher la prévision de latence pour le temps de réponse E/S sur la page analyse des charges de travail et sur la page Détails des événements. Lors de la collecte de cette activité, la prévision de latence 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 latence prévue 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 performance dynamique plus précis.

Au cours des 4 premiers jours qui suivent la surveillance d'une charge de travail par Unified Manager, si plus de 24 heures se sont écoulées depuis la dernière collecte de données, les graphiques de latence n'affichent pas les prévisions de latence pour cette charge de travail. Les événements détectés avant la dernière collection sont toujours disponibles.

> L'heure d'été change l'heure système, ce qui modifie la prévision de latence des statistiques de performances pour les charges de travail surveillées. Unified Manager commence immédiatement à corriger les prévisions de latence, ce qui prend environ 15 jours. Pendant cette période, vous pouvez continuer à utiliser Unified Manager, mais, puisque Unified Manager utilise les prévisions de latence pour détecter des événements dynamiques, certains événements peuvent ne pas être précis. Les événements détectés avant le changement de temps ne sont pas affectés.

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, une LUN, un partage NFS, un partage SMB/CIFS et un workload sont 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 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 impliquées dans l’efficacité du stockage, la réplication des données et l’état du système, notamment :

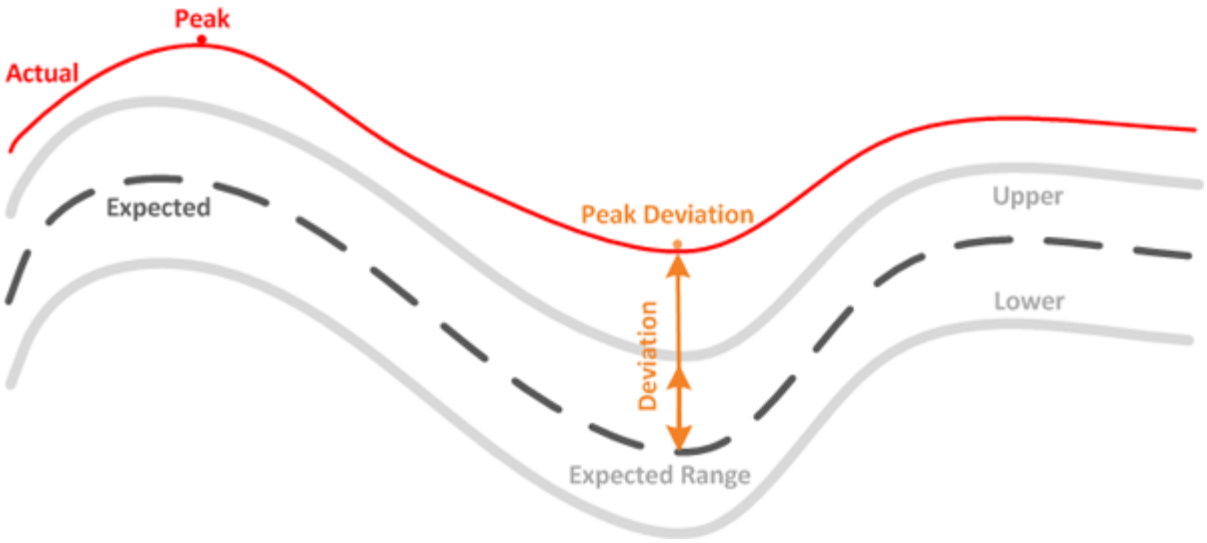
  - Efficacité du stockage, comme la déduplication
  - État de santé du disque, qui inclut la reconstruction RAID, le nettoyage du disque, etc
  - Réplication des données, notamment les copies SnapMirror
  - Activités de gestion
  - État de santé du système de fichiers, qui inclut les diverses activités 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.

Valeurs de mesure des performances des charges de travail

Unified Manager mesure les performances des charges de travail sur un cluster en fonction des valeurs statistiques historiques et attendues, qui constituent la prévision de latence des valeurs des workloads. Il compare les valeurs statistiques réelles de la charge de travail à la prévision de latence pour déterminer si les performances de la charge de travail sont trop élevées ou trop faibles. Un workload qui n’exécute pas comme prévu entraîne un événement de performance dynamique 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 performance, qui correspond aux limites supérieures de la prévision de latence. Le pic est la valeur réelle la plus élevée dans la période. L’écart mesure le changement entre les valeurs attendues (la prévision) 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	Pourcentage de la limite de qualité de service utilisée par les workloads dans le groupe de règles. 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 %.
Réel	La valeur des performances mesurée à un moment spécifique pour une charge de travail donnée.
Déviation	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. 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.

Mesure	Description
Attendu	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 attendue (prévision de latence) de valeurs.
Prévision de latence (plage prévue)	La prévision de latence est une prévision des valeurs de performance supérieure et inférieure attendues. 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 un événement de performance dynamique.
Pic	Valeur maximale mesurée sur une période de temps.
Déviation de crête	Valeur de déviation maximale mesurée sur une période de temps.
Profondeur de la file d'attente	Nombre de demandes d'E/S en attente du composant d'interconnexion.
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 (Mo/s), allant des charges de travail sur un cluster local au cluster partenaire dans une configuration MetroCluster.

Parmi les performances attendues

La prévision de latence est une prévision des valeurs de performance supérieure et inférieure attendues. 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 un événement de performance dynamique.

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 la plage d'activité attendue (prévision de latence), 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.

Mode de création de la prévision de latence

Unified Manager doit collecter au moins 3 jours d'activité de la charge de travail avant de commencer son analyse. Il est alors possible avant que la prévision de latence pour le temps de réponse d'E/S ne soit affichée 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. Après la collecte des 3 premiers jours d'activité, Unified Manager ajuste la latence prévue toutes les 24 heures à 12:00 afin de refléter les modifications apportées à l'activité des charges de travail et d'établir un seuil de performances dynamique plus précis.

L'heure d'été change l'heure système, ce qui modifie la prévision de latence des statistiques de performances pour les charges de travail surveillées. Unified Manager commence immédiatement à corriger les prévisions de latence, ce qui prend environ 15 jours. Pendant cette période, vous pouvez continuer à utiliser Unified Manager, mais, puisque Unified Manager utilise les prévisions de latence pour détecter des événements dynamiques, certains événements peuvent ne pas être précis. Les événements détectés avant le changement de temps ne sont pas affectés.

Mode d'utilisation de la prévision de latence dans l'analyse des performances

Unified Manager utilise les prévisions de latence pour représenter l'activité d'E/S type (temps de réponse) typique de vos charges de travail surveillées. Il vous alerte lorsque la latence réelle d'une charge de travail se situe au-dessus des limites supérieures de la prévision de latence, ce qui déclenche un événement de performance dynamique. Vous pouvez ainsi analyser le problème de performance et prendre des mesures correctives pour le résoudre.

La prévision de latence définit la base de 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 dynamique. Unified Manager utilise le modèle de base pour déterminer quand la latence réelle est au-dessus ou en dessous d'un seuil ou en dehors des limites de la 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 dynamique, 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 de la charge de travail élevée reste cohérente pendant une longue période, notamment plusieurs heures, Unified Manager considère que l'activité est normale et ajuste de manière dynamique les prévisions de latence afin de constituer le nouveau seuil de performance dynamique.

Certaines charges de travail peuvent avoir une activité faible et cohérente, où la latence prévue n'est pas fortement modifiée dans le temps. Pour minimiser le nombre 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 les latences sont beaucoup plus élevées que prévu.

Dans cet exemple, la latence d'un volume a une prévision, en gris, de 3.5 millisecondes par opération (ms/op) à sa plus faible et de 5.5 ms/opération à sa plus élevée. Si la latence réelle, en bleu, augmente soudainement à 10 ms/interruption, en raison d'un pic intermittent du trafic réseau ou d'un conflit sur un composant du cluster, il est alors au-dessus des prévisions de latence et a dépassé le seuil de performance dynamique.

Lorsque le trafic réseau a diminué ou que le composant de cluster n'est plus conflit, la latence est renvoyée dans la prévision de latence. 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

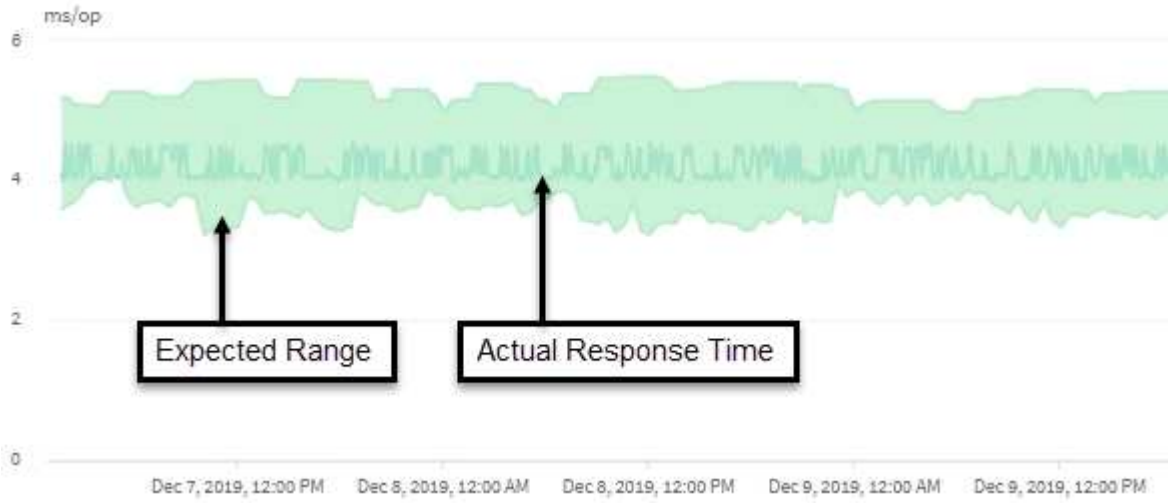
La latence (temps de réponse) correspond au temps nécessaire pour qu'un volume d'un cluster réponde aux demandes d'E/S des applications client. Unified Manager utilise la latence pour détecter les événements de performance et vous alerter.

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 dynamiques de la plage attendue (latence prévue), Unified Manager analyse l'événement afin de déterminer s'il s'agit d'un événement de performances que vous devrez résoudre. La latence est mesurée en millisecondes par opération (ms/op).

Dans le graphique Total de latence de la page analyse de charge de travail, vous pouvez visualiser une analyse des statistiques de latence afin de voir comment l'activité de processus individuels, tels que les requêtes de lecture et d'écriture, est comparé aux statistiques de latence globale. 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.



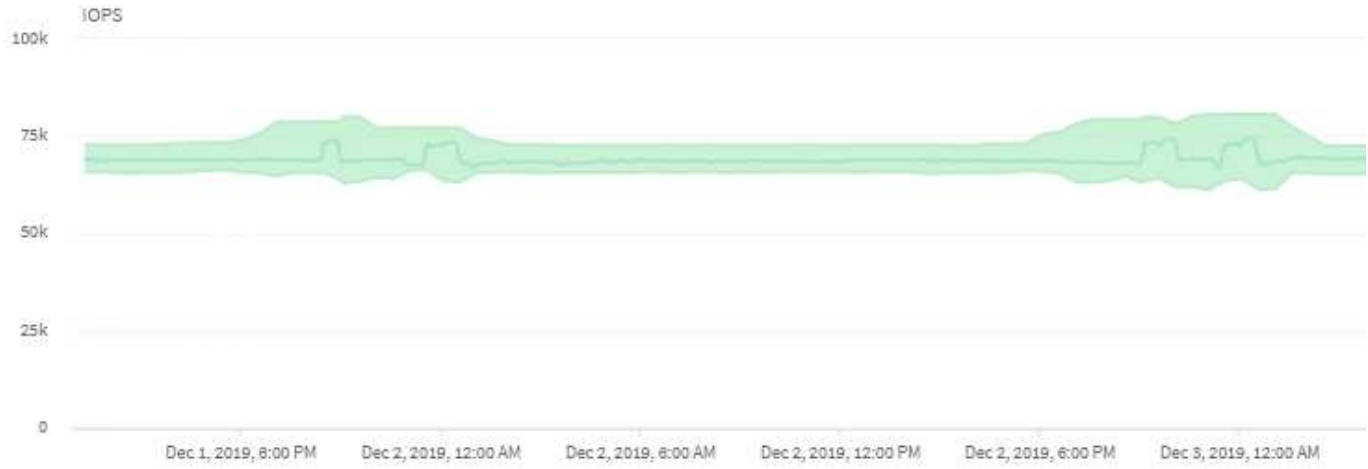
Cet exemple montre le graphique latence . L'activité du temps de réponse réel (latence) est une ligne bleue et la prévision de latence (plage prévue) est verte.

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).



L'exemple montre le graphique IOPS. Les statistiques d'opérations réelles sont une ligne bleue et la prévision des statistiques d'opérations d'E/S par seconde est verte.

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.

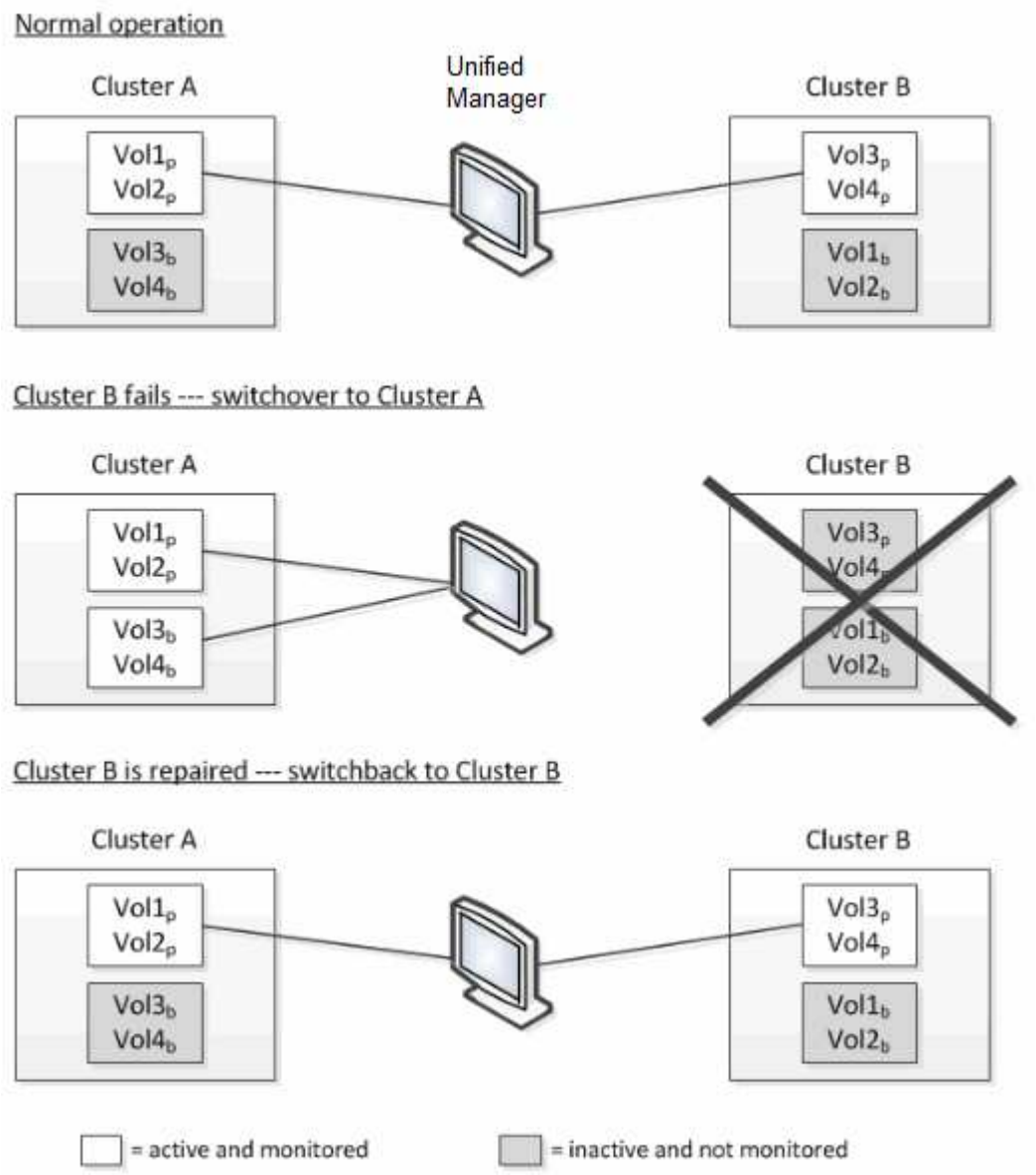
Unified Manager traite les clusters dans une configuration MetroCluster comme des clusters individuels. Il ne fait aucune distinction entre les clusters qui sont des partenaires ou établit un lien entre le débit d'écriture de chaque cluster.

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 :



Les volumes portant la référence p indiquent les volumes primaires, et les volumes dont l'noms 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 :

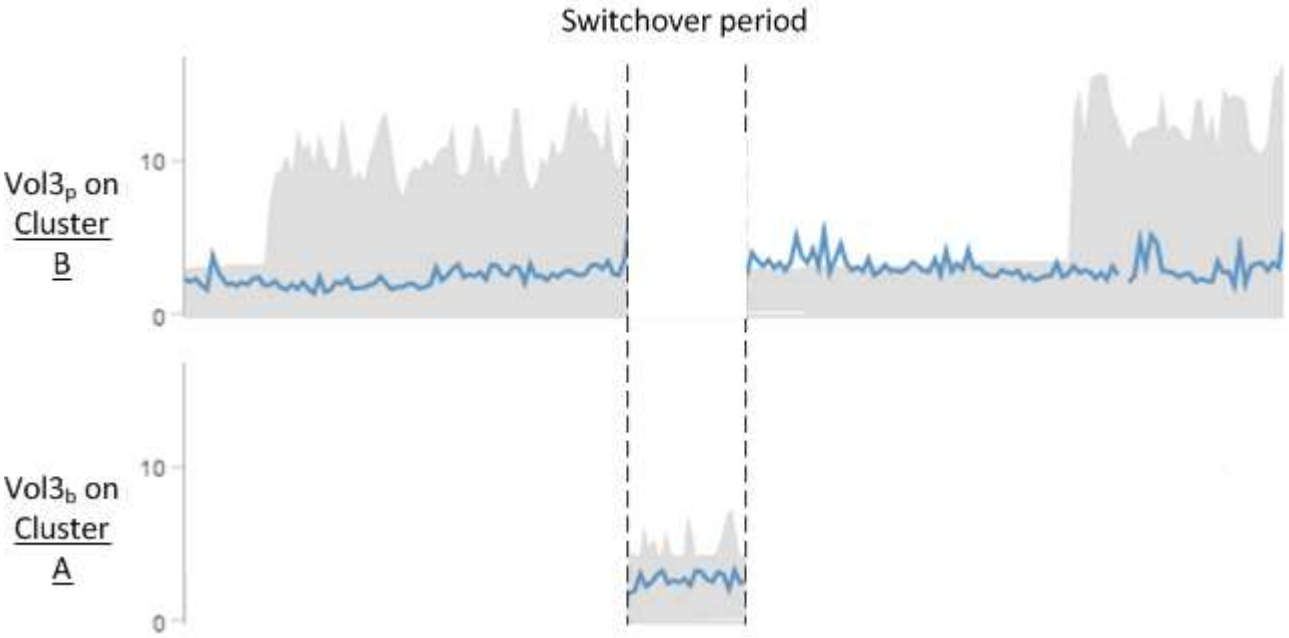
- Cluster A contient quatre volumes actifs : Vol1p, Vol2p, Vol3b et Vol4b.
- Le cluster B a quatre volumes inactifs : Vol3p, Vol4p, Vol1b et Vol2b.

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 » renterait 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.

En quoi sont les événements de performances

Les événements de performance sont des incidents liés aux performances des charges de travail sur un cluster. Ils vous aident à identifier les workloads avec des temps de réponse lents. Avec les événements de santé qui se sont produits en même temps, vous pouvez déterminer les problèmes qui pourraient avoir causé, ou contribué à, les délais de réponse lents.

Lorsque Unified Manager détecte plusieurs occurrences de la même condition d'événement pour le même composant de cluster, il traite toutes les occurrences comme un événement unique et non comme des événements distincts.

Analyse et notification des événements de performance

Les événements de performance vous signalent les problèmes de performances d'E/S concernant une charge de travail générée par des conflits sur 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 dynamique, Unified Manager déclenche un événement de performance afin de vous en avertir.

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 dynamique, qui est la limite supérieure de la prévision de latence :
  - Pour les volumes des agrégats HDD ou Flash Pool (hybride) (niveau local), les événements sont déclenchés uniquement lorsque la latence est supérieure à 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 (niveau cloud), 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.

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**

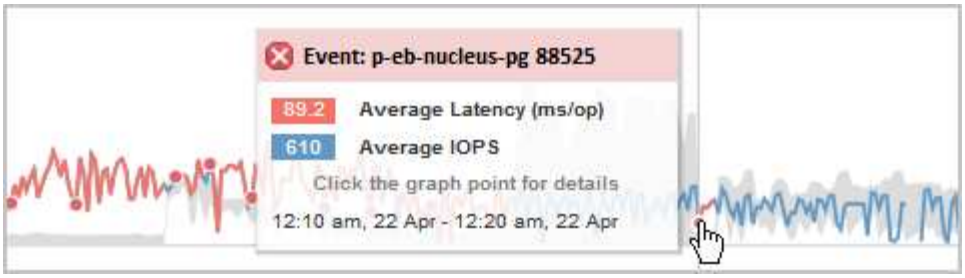
  Indique que l'événement de performance est actuellement actif (nouveau ou reconnu). Le problème à l'origine de l'incident n'a pas été corrigé lui-même ou n'a pas été résolu. Le compteur de performances de l'objet de stockage reste au-dessus du seuil de performance.

- **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 événements sont affichés sur la page Tableau de bord et sur de nombreuses autres pages de l'interface utilisateur, et les alertes pour ces événements 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 et sur la page analyse des charges de travail.



Dans cet exemple, un événement est indiqué par un point rouge (🔴) Sur le tableau latence. 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 Détails de l'événement et sur la page analyse de la charge de travail, vous pouvez interagir avec les événements de la manière suivante :

- Le déplacement de la souris sur un événement affiche un message indiquant l'ID de l'événement ainsi que la date et l'heure de détection de l'événement.

  S'il y a plusieurs événements pour la même période, le message indique le nombre d'événements.

- 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.

  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.

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 (prévision de latence). 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. Les prévisions de latence pour la charge de travail A sont de 5-10 millisecondes par opération (ms/op) et sa latence réelle est généralement d'environ 7 ms/op. La prévision de latence 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 de conflits sur le cluster, la latence des deux charges de travail augmente à 40 ms/opération, franchissement du seuil de performance dynamique, qui correspond aux limites supérieures des prévisions de latence et au déclenchement d'événements. 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 (🔴) indiquez 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 qui utilisent le ralentissement du composant 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, différentes étapes sont possibles : déplacer une ou plusieurs charges de travail vers un agrégat ou un nœud moins occupé, par exemple, afin de réduire les besoins globaux de la charge de travail sur l'agrégat en cours. 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 réseau externes sur le cluster. Le temps d'attente est le temps passé à attendre la fin des transactions « de transfert prêt » avant que le cluster puisse répondre à une demande d'E/S. Si le composant réseau constitue un conflit, cela signifie qu'un temps d'attente élevé au niveau de la couche de protocole 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.

Lors de l'utilisation d'un cluster All SAN Array dans une configuration active/active, la valeur de latence de traitement réseau s'affiche pour les deux nœuds afin que vous puissiez vérifier que les nœuds partagent la charge de manière égale.

- **Limite de qualité de service max**

  Représente le paramètre de débit maximal (crête) du groupe de règles de qualité de service (QoS) de stockage affecté à la charge de travail. 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.

- **Limite de qualité de service min**

  Représente la latence pour une charge de travail générée par le paramètre de débit de QoS minimal (attendu) attribué à d'autres workloads. Si, pour certaines charges de travail, la qualité de service minimale est définie sur la majorité de la bande passante pour garantir le débit promis, d'autres charges de travail sont restreintes et affichent une latence plus élevée.

- * 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.

- **Activation du volume**

  Processus permettant de suivre l'utilisation de tous les volumes actifs. Dans les environnements de grande taille où plus de 1000 volumes sont actifs, ce processus surveille en même temps le nombre de volumes stratégiques devant accéder aux ressources par le biais du nœud. Lorsque le nombre de volumes actifs simultanés dépasse le seuil maximal recommandé, certains volumes non critiques subissent une latence telle qu'elle est identifiée ici.

- **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 sa prévision de latence (plage 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.

Rôle	Description
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 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 le 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.