Collecte des données et contrôle des performances des workloads

Active IQ Unified Manager 9.13

NetApp
December 18, 2023

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. (Les volumes DP sont surveillés pour le trafic généré par les utilisateurs.)

- 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 :

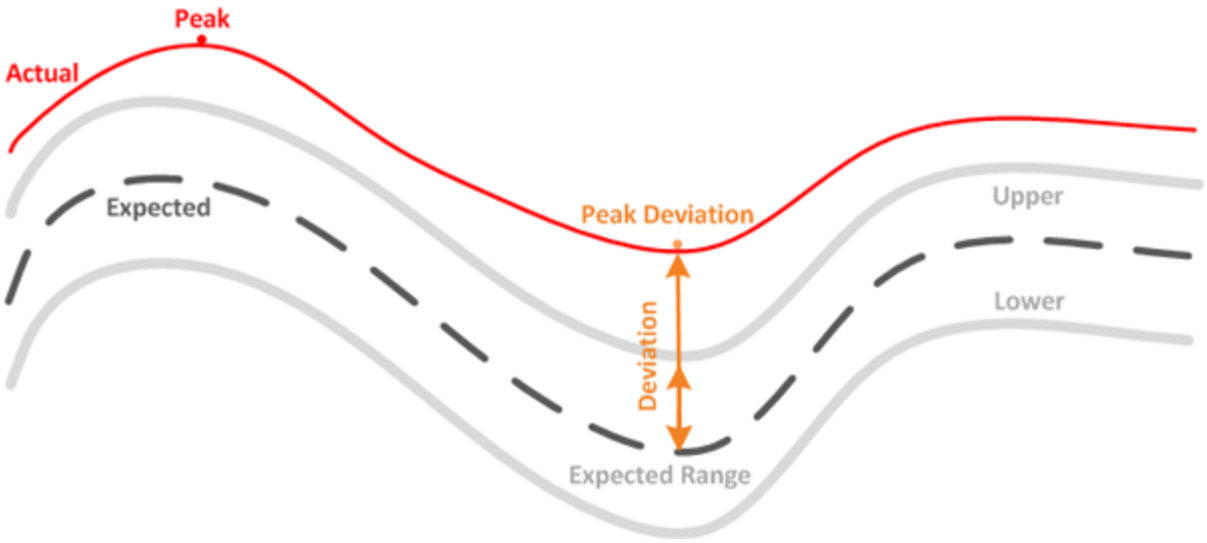
  - Efficacité du stockage, comme la déduplication
  - État de santé du disque, qui inclut la reconstruction RAID, le nettoyage du disque, etc
  - Réplication des données, notamment les copies SnapMirror
  - Activités de gestion
  - État de santé du système de fichiers, qui inclut les diverses activités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

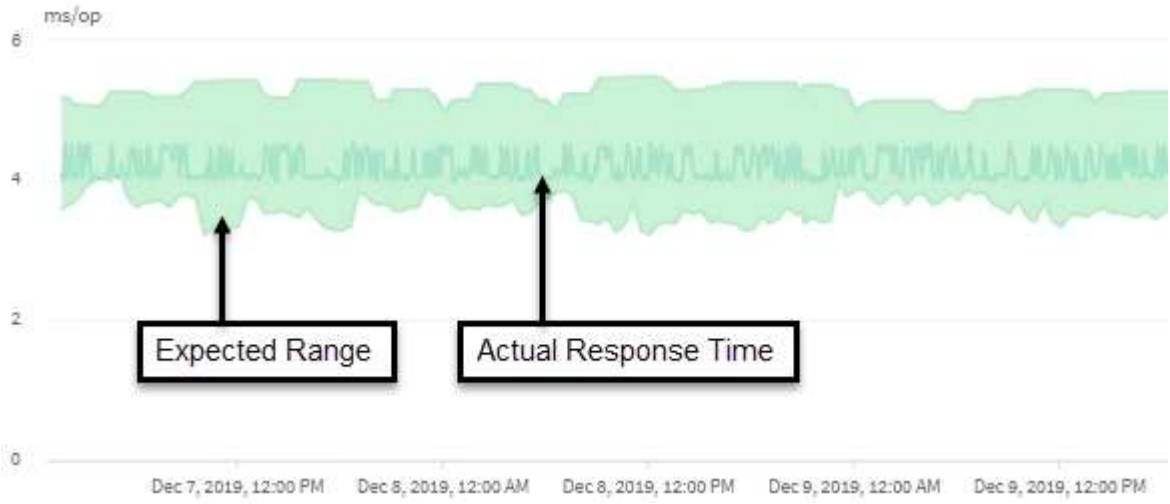
La latence (temps de réponse) correspond au temps nécessaire pour qu'un volume d'un cluster réponde aux demandes d'E/S des applications client. Unified Manager utilise la latence pour détecter les événements de performance et vous alerter.

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.



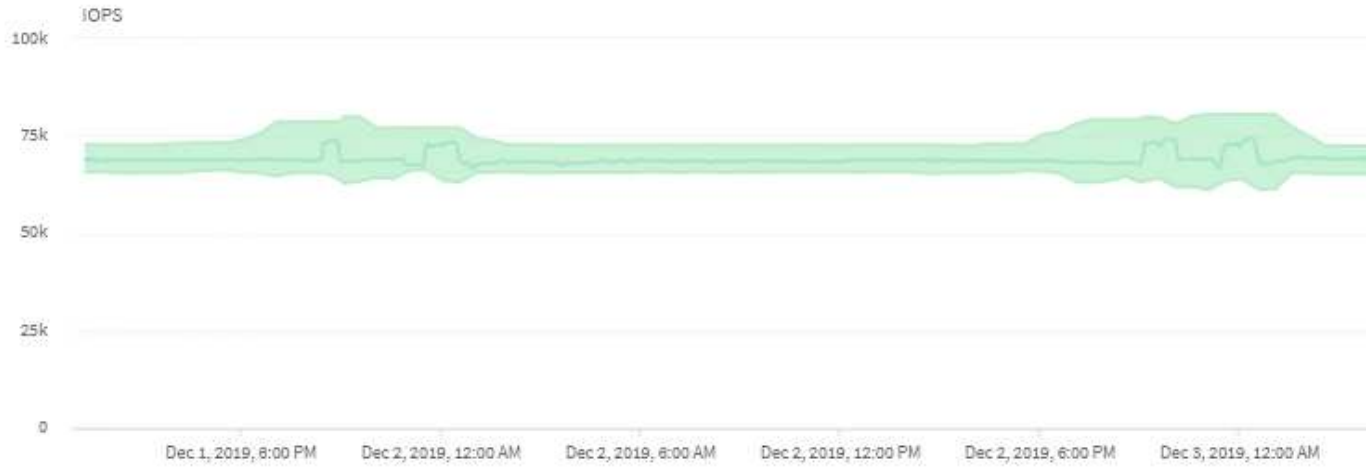
Cet exemple montre le graphique latence . L'activité du temps de réponse réel (latence) est une ligne bleue et la prévision de latence (plage prévue) est verte.

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.

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.

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.

Vous pouvez afficher le débit de tous les clusters d'une configuration MetroCluster en analysant les charges de travail des LUN et volumes correspondants à partir des écrans suivants. Vous pouvez filtrer les résultats par cluster. Dans le volet de navigation de gauche :

- **Stockage > clusters > Performance : vue tous les clusters**. Voir

- **Stockage > volumes > performances : vue tous les volumes**.
- **Stockage > LUN > performances : vue toutes les LUN**.
- **Analyse de la charge de travail > toutes les charges de travail**

Informations connexes

"Analyse et notification des événements de performance"

"Analyse des événements de performances pour une configuration MetroCluster"

"Rôles des charges de travail impliquées dans un événement de performance"

"L'identification des charges de travail victimes impliquées dans la mise en œuvre d'un événement de performance"

"L'identification des workloads dominants impliqués dans un événement de performance"

"L'identification des charges de travail Shark impliquées dans un événement de performance"