Coleta de dados e monitoramento do desempenho da carga de trabalho

Active IQ Unified Manager 9.8

NetApp
January 31, 2025

Índice

Coleta de dados e monitoramento do desempenho da carga de trabalho

O Unified Manager coleta e analisa a atividade do workload a cada 5 minutos para identificar eventos de performance e detecta alterações de configuração a cada 15 minutos. Ele retém um máximo de 30 dias de dados de eventos e performance históricos de 5 minutos e usa esses dados para prever o intervalo de latência esperado para todos os workloads monitorados.

O Unified Manager deve coletar no mínimo 3 dias de atividade do workload antes que ele possa começar a análise e antes que a previsão de latência do tempo de resposta de e/S possa ser exibida na página análise do workload e na página de detalhes do evento. Enquanto essa atividade está sendo coletada, a previsão de latência não exibe todas as alterações que ocorrem na atividade da carga de trabalho. Após coletar 3 dias de atividade, o Unified Manager ajusta a previsão de latência a cada 24 horas às 12:00 da manhã, para refletir as alterações nas atividades do workload e estabelecer um limite de performance dinâmico mais preciso.

Durante os primeiros 4 dias em que o Unified Manager está monitorando uma carga de trabalho, se mais de 24 horas passaram desde a última coleta de dados, os gráficos de latência não exibirão a previsão de latência para essa carga de trabalho. Os eventos detetados antes da última coleção ainda estão disponíveis.

> O horário de verão (DST) altera a hora do sistema, o que altera a previsão de latência das estatísticas de desempenho para cargas de trabalho monitoradas. O Unified Manager começa imediatamente a corrigir a previsão de latência, que leva aproximadamente 15 dias para ser concluída. Durante esse período, você pode continuar usando o Unified Manager, mas, como o Unified Manager usa a previsão de latência para detetar eventos dinâmicos, alguns eventos podem não ser precisos. Os eventos detetados antes da alteração de hora não são afetados.

Tipos de workloads monitorados pelo Unified Manager

Você pode usar o Unified Manager para monitorar a performance de dois tipos de workloads: Definido pelo usuário e definido pelo sistema.

- ***cargas de trabalho definidas pelo usuário***

  A taxa de transferência de e/S das aplicações para o cluster. Estes são processos envolvidos em pedidos de leitura e escrita. Um volume, LUN, compartilhamento NFS, compartilhamento SMB/CIFS e um workload são um workload definido pelo usuário.

  > O Unified Manager monitora apenas a atividade do workload no cluster. Ele não monitora os aplicativos, os clientes ou os caminhos entre os aplicativos e o cluster.

  Se uma ou mais das opções a seguir for verdadeira para uma carga de trabalho, ela não poderá ser monitorada pelo Unified Manager:

  - É uma cópia de proteção de dados (DP) no modo somente leitura. (Os volumes DP são monitorados quanto ao tráfego gerado pelo usuário.)
  - É um clone de dados off-line.
  - É um volume espelhado em uma configuração do MetroCluster.

- ***cargas de trabalho definidas pelo sistema***

  Os processos internos envolvidos com eficiência de storage, replicação de dados e integridade do sistema, incluindo:

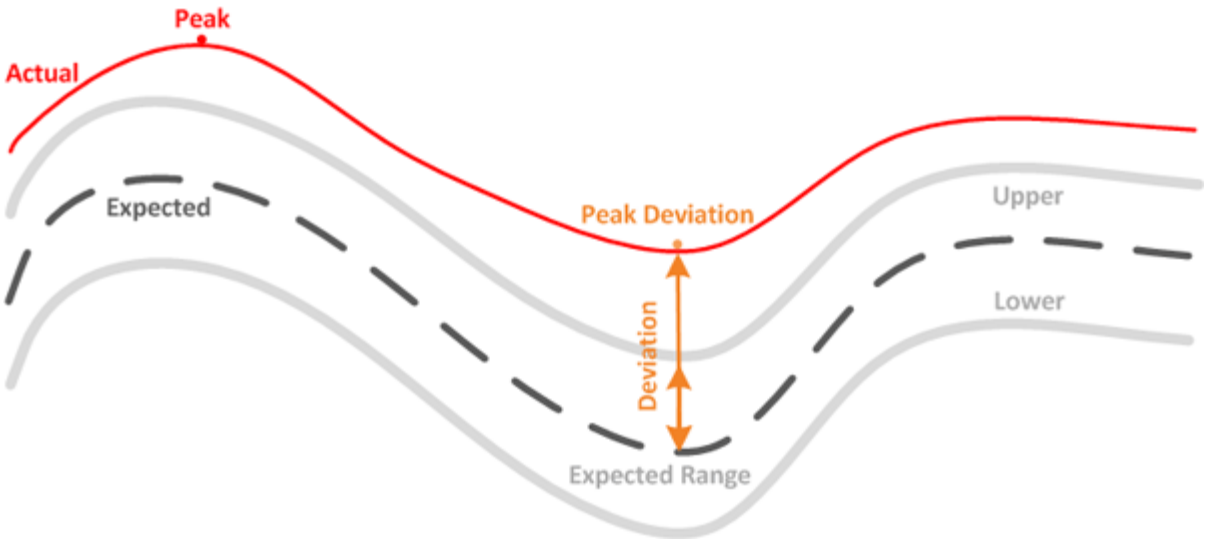
  - Eficiência de storage, como deduplicação
  - Integridade do disco, que inclui RAID Reconstruct, análise de disco e assim por diante
  - Replicação de dados, como cópias SnapMirror
  - Atividades de gestão
  - Integridade do sistema de arquivos, que inclui várias atividades do WAFL
  - Scanners de sistema de arquivos, como WAFL scan
  - Descarga de cópia, como operações de eficiência de storage descarregadas de hosts VMware
  - Integridade do sistema, como movimentos de volume, compactação de dados etc.
  - Volumes não monitorizados

Os dados de performance para workloads definidos pelo sistema são exibidos na GUI somente quando o componente de cluster usado por esses workloads está na contenção. Por exemplo, você não pode pesquisar o nome de uma carga de trabalho definida pelo sistema para exibir seus dados de performance na GUI.

Valores de medição de performance de workload

O Unified Manager mede o desempenho de workloads em um cluster com base em valores estatísticos históricos e esperados, que formam a previsão de latência de valores para as cargas de trabalho. Ele compara os valores estatísticos reais de workload com a previsão de latência para determinar quando a performance do workload é muito alta ou muito baixa. Uma carga de trabalho que não está funcionando como esperado aciona um evento de desempenho dinâmico para notificá-lo.

Na ilustração a seguir, o valor real, em vermelho, representa as estatísticas reais de desempenho no período de tempo. O valor real cruzou o limite de desempenho, que é os limites superiores da previsão de latência. O pico é o valor real mais alto no período de tempo. O desvio mede a mudança entre os valores esperados (a previsão) e os valores reais, enquanto o desvio de pico indica a maior mudança entre os valores esperados e os valores reais.



A tabela a seguir lista os valores de medição de desempenho da carga de trabalho.

Medição	Descrição
Atividade	Porcentagem do limite de QoS usado pelos workloads no grupo de políticas. Se o Unified Manager detetar uma alteração em um grupo de políticas, como adicionar ou remover um volume ou alterar o limite de QoS, os valores real e esperado poderão exceder 100% do limite definido. Se um valor exceder 100% do limite definido, é apresentado como >100%. Se um valor for inferior a 1% do limite definido, é apresentado como inferior a 1%.
Real	O valor de desempenho medido em um momento específico para uma determinada carga de trabalho.
Desvio	A mudança entre os valores esperados e os valores reais. É a relação do valor real menos o valor esperado para o valor superior do intervalo esperado menos o valor esperado. Um valor de desvio negativo indica que o desempenho da carga de trabalho é inferior ao esperado, enquanto um valor de desvio positivo indica que o desempenho da carga de trabalho é superior ao esperado.

Medição	Descrição
Esperado	Os valores esperados são baseados na análise de dados históricos de performance para uma determinada carga de trabalho. O Unified Manager analisa esses valores estatísticos para determinar o intervalo esperado (previsão de latência) dos valores.
Previsão de latência (intervalo esperado)	A previsão de latência é uma previsão do que os valores de desempenho superior e inferior devem ser em um momento específico. Para a latência do workload, os valores superiores formam o limite de performance. Quando o valor real cruza o limite de performance, o Unified Manager aciona um evento de performance dinâmico.
Pico	O valor máximo medido durante um período de tempo.
Desvio máximo	O valor de desvio máximo medido durante um período de tempo.
Profundidade da fila	O número de solicitações de e/S pendentes que estão aguardando no componente de interconexão.
Utilização	Para os componentes de processamento de rede, Data Processing e agregado, a porcentagem de tempo de ocupado para concluir as operações de carga de trabalho por um período de tempo. Por exemplo, a porcentagem de tempo para o processamento de rede ou os componentes do Data Processing processarem uma solicitação de e/S ou para um agregado atender a uma solicitação de leitura ou gravação.
Taxa de transferência de gravação	A quantidade de taxa de transferência de gravação, em megabytes por segundo (MB/s), desde cargas de trabalho em um cluster local até o cluster de parceiros em uma configuração do MetroCluster.

Qual é a faixa de desempenho esperada

A previsão de latência é uma previsão do que os valores de desempenho superior e inferior devem ser em um momento específico. Para a latência do workload, os valores superiores formam o limite de performance. Quando o valor real cruza o limite de performance, o Unified Manager aciona um evento de performance dinâmico.

Por exemplo, durante o horário comercial regular entre as 9:00h e as 5:00h, a maioria dos funcionários pode verificar seu e-mail entre as 9:00h e as 10:30H. o aumento da demanda nos servidores de e-mail significa um

aumento na atividade de carga de trabalho no armazenamento de back-end durante esse período. Os funcionários podem notar um tempo de resposta lento de seus clientes de e-mail.

Durante a hora de almoço entre as 12:00h e as 1:00h e no final do dia de trabalho após as 5:00h, a maioria dos funcionários provavelmente está longe de seus computadores. A demanda nos servidores de e-mail geralmente diminui, também diminuindo a demanda no armazenamento de back-end. Como alternativa, pode haver operações de carga de trabalho agendadas, como backups de armazenamento ou verificação de vírus, que começam após as 5:00 horas e aumentam a atividade no armazenamento de back-end.

Ao longo de vários dias, o aumento e a diminuição da atividade de workload determinam o intervalo esperado (previsão de latência) da atividade, com limites superior e inferior para uma carga de trabalho. Quando a atividade de carga de trabalho real para um objeto está fora dos limites superior ou inferior e permanece fora dos limites por um período de tempo, isso pode indicar que o objeto está sendo usado em excesso ou subutilizado.

Como a previsão de latência é formada

O Unified Manager deve coletar no mínimo 3 dias de atividade do workload antes que ele possa iniciar a análise e antes que a previsão de latência do tempo de resposta de e/S possa ser exibida na GUI. A coleta de dados mínima necessária não é responsável por todas as alterações que ocorrem na atividade da carga de trabalho. Após coletar os primeiros 3 dias de atividade, o Unified Manager ajusta a previsão de latência a cada 24 horas às 12:00 da manhã para refletir as alterações na atividade do workload e estabelecer um limite de performance dinâmico mais preciso.

O horário de verão (DST) altera a hora do sistema, o que altera a previsão de latência das estatísticas de desempenho para cargas de trabalho monitoradas. O Unified Manager começa imediatamente a corrigir a previsão de latência, que leva aproximadamente 15 dias para ser concluída. Durante esse período, você pode continuar usando o Unified Manager, mas, como o Unified Manager usa a previsão de latência para detetar eventos dinâmicos, alguns eventos podem não ser precisos. Os eventos detetados antes da alteração de hora não são afetados.

Como a previsão de latência é usada na análise de desempenho

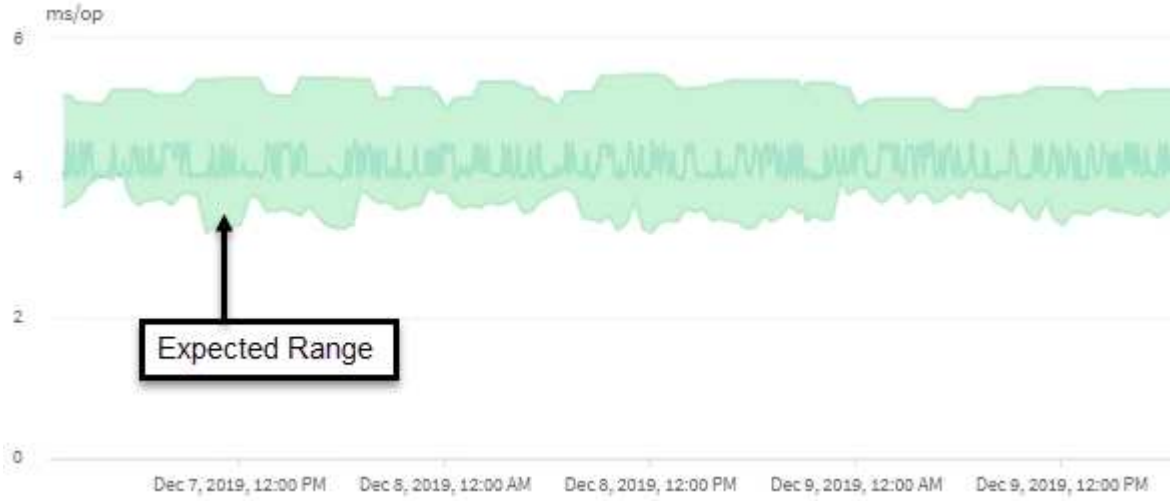
O Unified Manager usa a previsão de latência para representar a atividade típica de latência de e/S (tempo de resposta) dos workloads monitorados. Ele alerta quando a latência real de um workload está acima dos limites superiores da previsão de latência, que aciona um evento de performance dinâmico, para que você possa analisar o problema de performance e tomar medidas corretivas para resolvê-lo.

A previsão de latência define a linha de base de desempenho para a carga de trabalho. Com o tempo, o Unified Manager aprende com medições de desempenho anteriores para prever os níveis de desempenho e atividade esperados para a carga de trabalho. O limite superior do intervalo esperado estabelece o limite de desempenho dinâmico. O Unified Manager usa a linha de base para determinar quando a latência real está acima ou abaixo de um limite ou fora dos limites de seu intervalo esperado. A comparação entre os valores reais e os valores esperados cria um perfil de performance para a carga de trabalho.

Quando a latência real de um workload excede o limite dinâmico de performance, devido à contenção em um componente do cluster, a latência é alta e o workload opera mais lentamente do que o esperado. O desempenho de outras cargas de trabalho que compartilham os mesmos componentes de cluster também pode ser mais lento do que o esperado.

O Unified Manager analisa o evento de cruzamento de limites e determina se a atividade é um evento de desempenho. Se a atividade de alto workload permanecer consistente por um longo período de tempo, como várias horas, o Unified Manager considera a atividade normal e ajusta dinamicamente a previsão de latência para formar o novo limite dinâmico de performance.

Algumas cargas de trabalho podem ter atividades consistentemente baixas, em que a previsão de latência para latência não tem uma alta taxa de alteração ao longo do tempo. Para minimizar o número de eventos durante a análise de eventos de performance, o Unified Manager aciona um evento apenas para volumes de baixa atividade cujas operações e latências são muito maiores do que o esperado.



Neste exemplo, a latência de um volume tem uma previsão de latência, em cinza, de 3,5 milissegundos por operação (ms/op) no menor e 5,5 ms/op no máximo. Se a latência real, em azul, aumentar repentinamente para 10 ms/op, devido a um pico intermitente no tráfego de rede ou contenção em um componente de cluster, ela fica então acima da previsão de latência e excedeu o limite de desempenho dinâmico.

Quando o tráfego de rede diminuiu ou o componente do cluster não está mais na contenção, a latência retorna dentro da previsão de latência. Se a latência permanecer em ou acima de 10 ms/op por um longo período de tempo, talvez seja necessário tomar medidas corretivas para resolver o evento.

Como o Unified Manager usa a latência do workload para identificar problemas de performance

A latência do workload (tempo de resposta) é o tempo necessário para um volume em um cluster responder a solicitações de e/S de aplicativos clientes. O Unified Manager usa a latência para detectar e alertar você sobre eventos de performance.

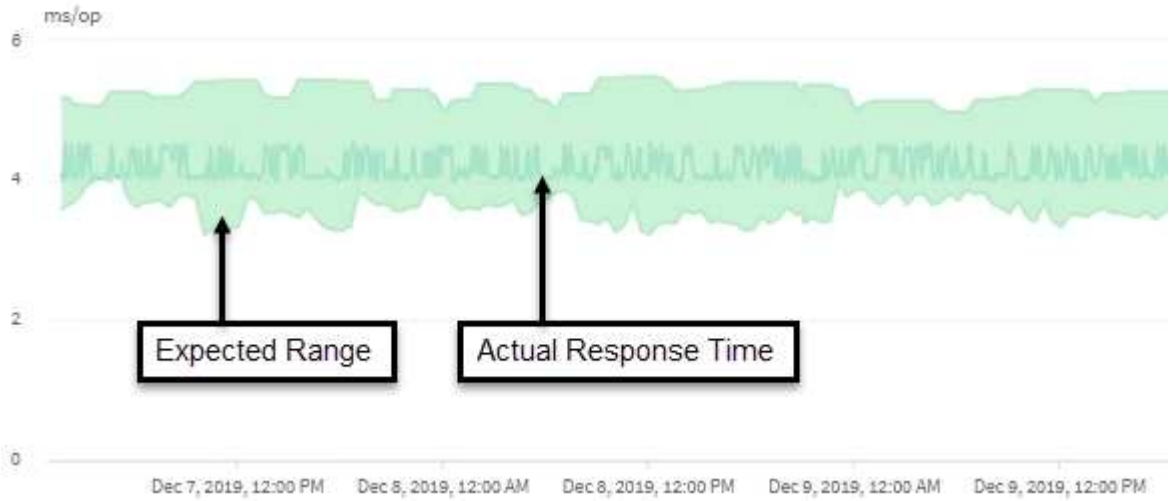
Uma alta latência significa que as solicitações de aplicativos para um volume em um cluster estão demorando mais do que o normal. A causa da alta latência pode estar no próprio cluster, devido à contenção em um ou mais componentes do cluster. A alta latência também pode ser causada por problemas fora do cluster, como gargalos de rede, problemas com o cliente que hospeda os aplicativos ou problemas com os próprios aplicativos.

O Unified Manager monitora apenas a atividade do workload no cluster. Ele não monitora os aplicativos, os clientes ou os caminhos entre os aplicativos e o cluster.

As operações no cluster, como fazer backups ou executar deduplicação, que aumentam a demanda por componentes de cluster compartilhados por outros workloads, também podem contribuir para a alta latência.

Se a latência real exceder o limite de desempenho dinâmico do intervalo esperado (previsão de latência), o Unified Manager analisa o evento para determinar se é um evento de desempenho que talvez você precise resolver. A latência é medida em milissegundos por operação (ms/op).

No gráfico total de latência na página análise de workload, é possível visualizar uma análise das estatísticas de latência para ver como a atividade de processos individuais, como solicitações de leitura e gravação, se compara às estatísticas de latência geral. A comparação ajuda você a determinar quais operações têm a atividade mais alta ou se operações específicas têm atividade anormal que está afetando a latência de um volume. Ao analisar eventos de desempenho, você pode usar as estatísticas de latência para determinar se um evento foi causado por um problema no cluster. Você também pode identificar as atividades específicas de workload ou os componentes de cluster envolvidos no evento.



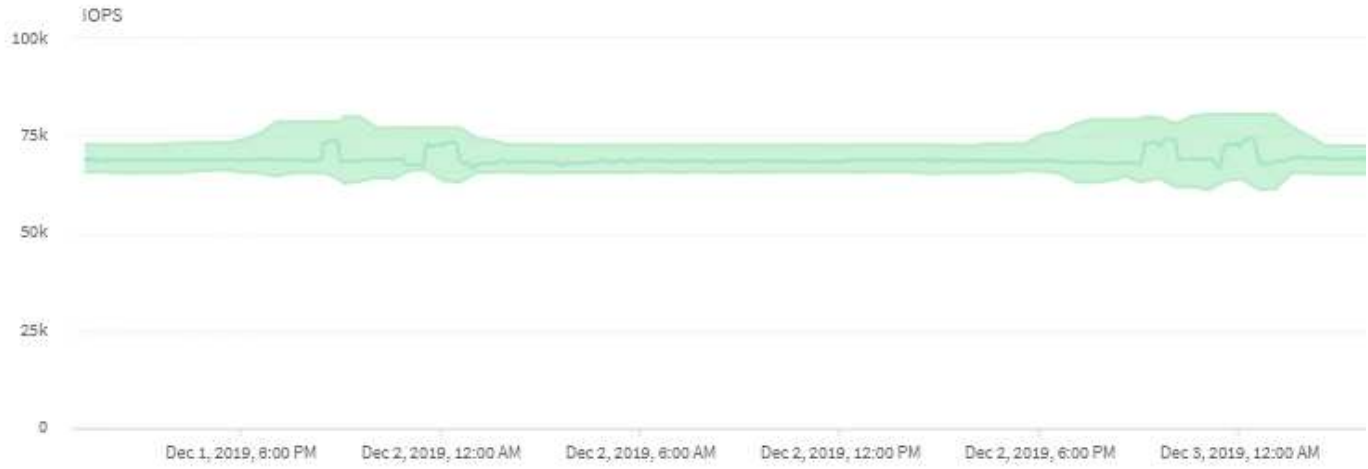
Este exemplo mostra o gráfico de latência . A atividade de tempo de resposta real (latência) é uma linha azul e a previsão de latência (intervalo esperado) é verde.

Pode haver lacunas na linha azul se o Unified Manager não conseguir coletar dados. Isso pode ocorrer porque o cluster ou o volume estava inalcançável, o Unified Manager foi desativado durante esse tempo ou a coleção demorava mais do que o período de coleta de 5 minutos.

Como as operações do cluster podem afetar a latência do workload

As operações (IOPS) representam a atividade de todos os workloads definidos pelo usuário e definidos pelo sistema em um cluster. As estatísticas de IOPS ajudam a determinar se os processos de cluster, como fazer backups ou executar deduplicação, estão impactando a latência do workload (tempo de resposta) ou podem ter causado ou contribuído para um evento de performance.

Ao analisar eventos de desempenho, você pode usar as estatísticas de IOPS para determinar se um evento de desempenho foi causado por um problema no cluster. Você pode identificar as atividades específicas de carga de trabalho que podem ter sido os principais contribuintes para o evento de performance. As operações de entrada/saída por segundo (operações/seg) são medidas em operações por segundo (operações/seg).



Este exemplo mostra o gráfico de IOPS. As estatísticas de operações reais são uma linha azul e a previsão de operações de IOPS é verde.

Em alguns casos em que um cluster está sobrecarregado, o Unified Manager pode exibir a mensagem `Data collection is taking too long on Cluster cluster_name`. Isso significa que não foram coletadas estatísticas suficientes para que o Unified Manager analise. Você precisa reduzir os recursos que o cluster está usando para que as estatísticas possam ser coletadas.

Monitoramento de desempenho das configurações do MetroCluster

Com o Unified Manager, você monitora a taxa de transferência de gravação entre clusters em uma configuração do MetroCluster para identificar workloads com uma taxa de transferência de gravação alta. Se esses workloads de alta performance fizerem com que outros volumes no cluster local tenham tempos de resposta de e/S altos, o Unified Manager acionará eventos de desempenho para notificá-lo.

Quando um cluster local em uma configuração do MetroCluster espelha seus dados em seu cluster de parceiros, os dados são gravados no NVRAM e transferidos pelos links de interswitch (ISLs) para os agregados remotos. O Unified Manager analisa o NVRAM para identificar workloads cuja alta taxa de transferência de gravação sobreutiliza o NVRAM, colocando o NVRAM na contenção.

Cargas de trabalho cujo desvio no tempo de resposta excedeu o limite de desempenho são chamadas *vítimas* e cargas de trabalho cujo desvio no throughput de gravação para o NVRAM é maior do que o habitual, causando a contenção, são chamadas *bullies*. Como apenas as solicitações de gravação são espelhadas no cluster de parceiros, o Unified Manager não analisa a taxa de transferência de leitura.

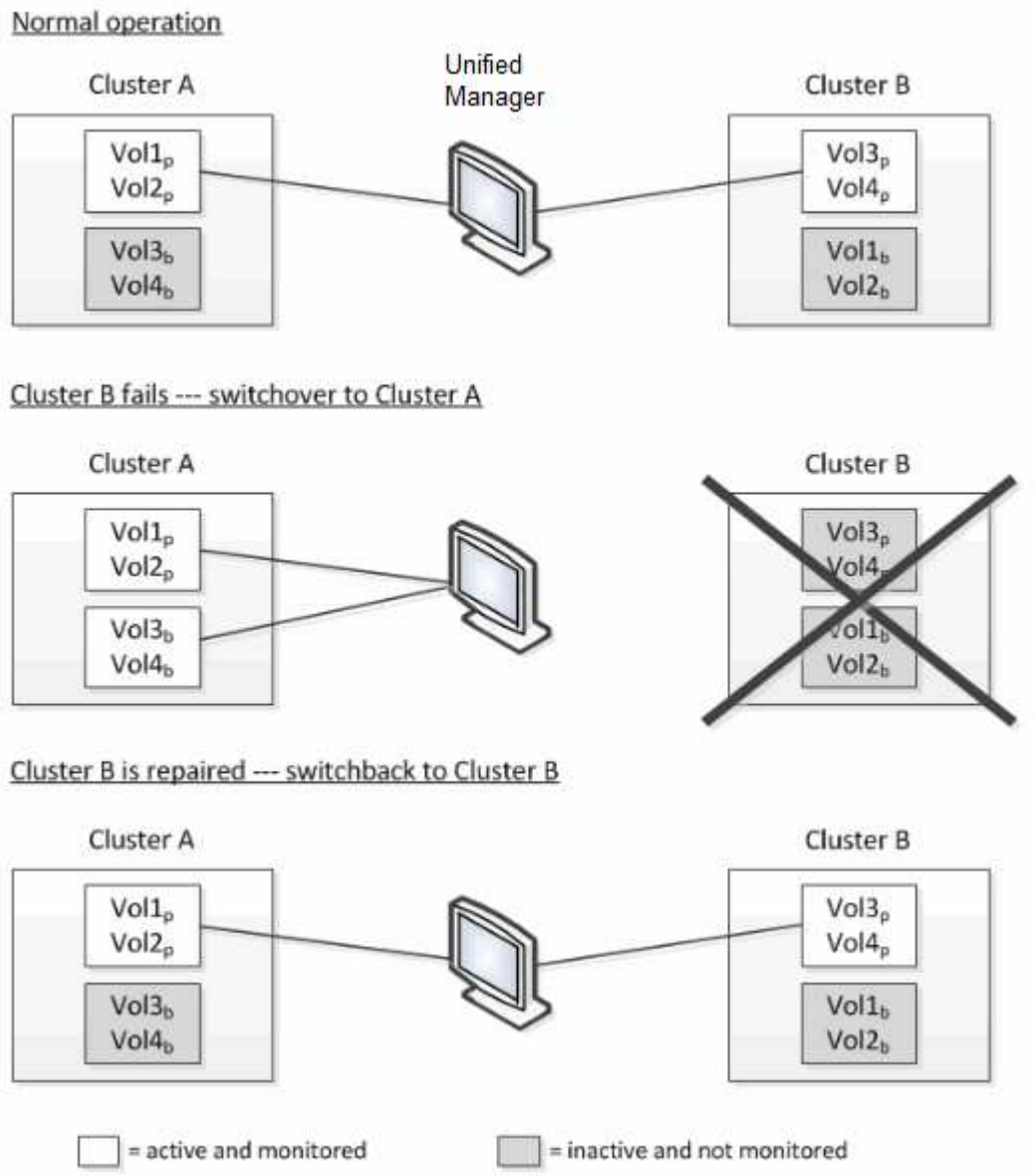
O Unified Manager trata os clusters em uma configuração do MetroCluster como clusters individuais. Isso não faz distinção entre clusters que são parceiros ou correlacionam a taxa de transferência de gravação de cada cluster.

Comportamento do volume durante o switchover e o switchback

Os eventos que acionam um switchover ou switchback fazem com que os volumes ativos sejam movidos de um cluster para o outro cluster no grupo de recuperação de desastres.

Os volumes no cluster que estavam ativos e fornecendo dados aos clientes são interrompidos, e os volumes no outro cluster são ativados e começam a fornecer dados. O Unified Manager monitora apenas os volumes ativos e em execução.

Como os volumes são movidos de um cluster para outro, é recomendável que você monitore os dois clusters. Uma única instância do Unified Manager pode monitorar ambos os clusters em uma configuração do MetroCluster, mas às vezes a distância entre os dois locais exige o uso de duas instâncias do Unified Manager para monitorar ambos os clusters. A figura a seguir mostra uma única instância do Unified Manager:



Os volumes com p em seus nomes indicam os volumes primários, e os volumes com b em seus nomes são volumes de backup espelhados criados pelo SnapMirror.

Durante o funcionamento normal:

- O cluster A tem dois volumes ativos: Vol1p e Vol2p.
- O cluster B tem dois volumes ativos: Vol3p e Vol4p.
- O cluster A tem dois volumes inativos: Vol3b e Vol4b.

- O cluster B tem dois volumes inativos: Vol1b e Vol2b.

As informações referentes a cada um dos volumes ativos (estatísticas, eventos etc.) são coletadas pelo Unified Manager. As estatísticas Vol1p e Vol2p são coletadas pelo Cluster A e as estatísticas Vol3p e Vol4p são coletadas pelo Cluster B.

Após uma falha catastrófica, causa um switchover de volumes ativos do cluster B para o cluster A:

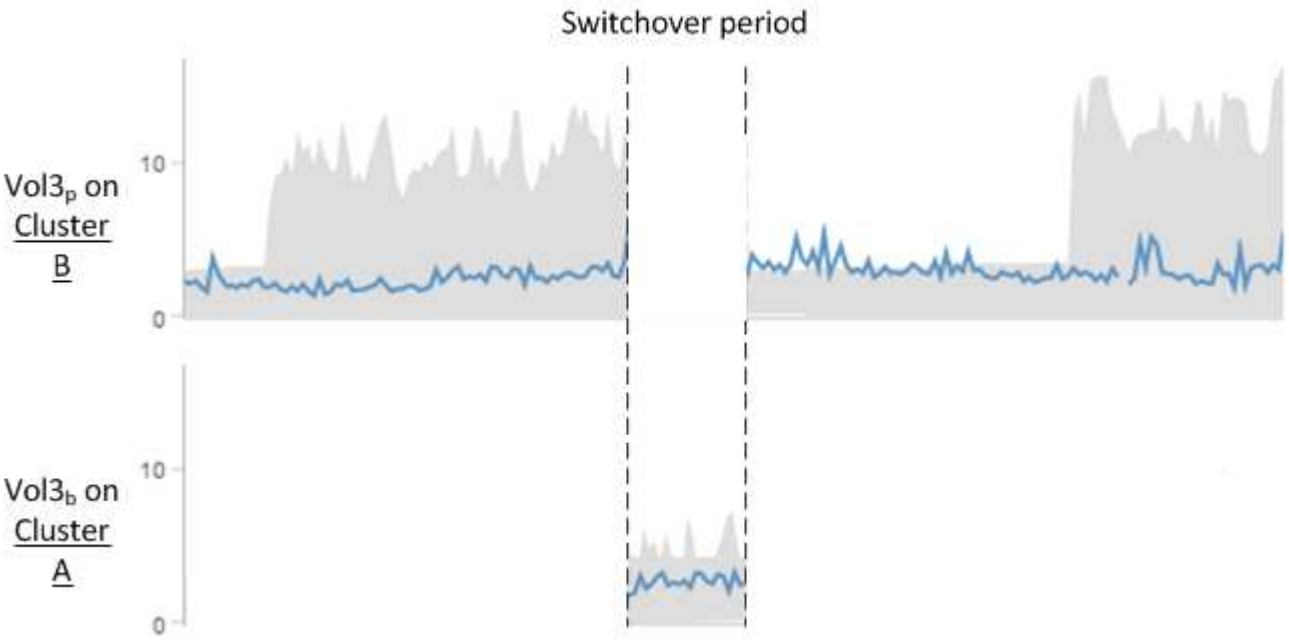
- O cluster A tem quatro volumes ativos: Vol1p, Vol2p, Vol3b e Vol4b.
- O cluster B tem quatro volumes inativos: Vol3p, Vol4p, Vol1b e Vol2b.

Como durante a operação normal, as informações referentes a cada um dos volumes ativos são coletadas pelo Unified Manager. Mas neste caso, as estatísticas Vol1p e Vol2p são coletadas pelo Cluster A, e as estatísticas Vol3b e Vol4b também são coletadas pelo Cluster A.

Observe que Vol3p e Vol3b não são os mesmos volumes, porque estão em clusters diferentes. As informações do Unified Manager para Vol3p não são as mesmas que Vol3b:

- Durante o switchover para o cluster A, as estatísticas e os eventos do Vol3p não são visíveis.
- Na primeira mudança, Vol3b parece um novo volume sem informações históricas.

Quando o cluster B é reparado e um switchback é executado, o Vol3p é novamente ativo no cluster B, com as estatísticas históricas e uma lacuna de estatísticas para o período durante o switchover. O Vol3b não pode ser visualizado a partir do cluster A até que ocorra outro switchover:



- Os volumes MetroCluster que estão inativos, por exemplo, Vol3b no cluster A após o switchback, são identificados com a mensagem ""este volume foi excluído"". O volume não é realmente excluído, mas não está sendo monitorado pelo Unified Manager, porque não é o volume ativo.
- Se um único Gerenciador unificado estiver monitorando ambos os clusters em uma configuração do MetroCluster, a pesquisa de volume retornará informações sobre o volume que estiver ativo naquele momento. Por exemplo, uma pesquisa por "'Vol3'" retornará estatísticas e eventos para Vol3b no Cluster A se um switchover tiver ocorrido e Vol3 se tornar ativo no Cluster A.

Quais são os eventos de desempenho

Os eventos de desempenho são incidentes relacionados ao desempenho da carga de trabalho em um cluster. Eles ajudam a identificar workloads com tempos de resposta lentos. Juntamente com eventos de saúde que ocorreram ao mesmo tempo, você pode determinar os problemas que podem ter causado ou contribuído para os tempos de resposta lentos.

Quando o Unified Manager deteta várias ocorrências da mesma condição de evento para o mesmo componente de cluster, ele trata todas as ocorrências como um único evento, não como eventos separados.

Análise e notificação de eventos de performance

Os eventos de desempenho notificam você sobre problemas de desempenho de e/S em uma carga de trabalho causada pela contenção em um componente do cluster. O Unified Manager analisa o evento para identificar todos os workloads envolvidos, o componente em contenção e se o evento ainda é um problema que talvez você precise resolver.

O Unified Manager monitora a latência de e/S (tempo de resposta) e IOPS (operações) de volumes em um cluster. Quando outras cargas de trabalho usam excessivamente um componente de cluster, por exemplo, o componente está na contenção e não pode ter desempenho em um nível ideal para atender às demandas de workload. O desempenho de outros workloads que estão usando o mesmo componente pode ser afetado, causando o aumento de suas latências. Se a latência ultrapassar o limite de desempenho dinâmico, o Unified Manager acionará um evento de desempenho para notificá-lo.

Análise de eventos

O Unified Manager realiza as seguintes análises, usando as estatísticas de desempenho dos 15 dias anteriores, para identificar os workloads da vítima, os workloads bully e o componente do cluster envolvido em um evento:

- Identifica cargas de trabalho da vítima cuja latência ultrapassou o limite de desempenho dinâmico, que é o limite superior da previsão de latência:
  - Para volumes em agregados híbridos HDD ou Flash Pool (camada local), os eventos são acionados somente quando a latência é maior que 5 milissegundos (ms) e o IOPS é mais de 10 operações por segundo (operações/seg).
  - Para volumes em agregados all-SSD ou agregados FabricPool (camada de nuvem), os eventos são acionados apenas quando a latência é superior a 1 ms e o IOPS é superior a 100 operações/seg
- Identifica o componente do cluster na contenção.

Se a latência das cargas de trabalho da vítima na interconexão de cluster for superior a 1 ms, o Unified Manager tratará isso como significativo e acionará um evento para a interconexão de cluster.

- Identifica as cargas de trabalho bully que estão sobreusando o componente do cluster e fazendo com que ele esteja na contenção.
- Classifica as cargas de trabalho envolvidas, com base em seu desvio na utilização ou atividade de um componente de cluster, para determinar quais bullies têm a maior alteração no uso do componente de cluster e quais vítimas são as mais impactadas.

Um evento pode ocorrer por apenas um breve momento e depois se corrigir depois que o componente que está usando não está mais em disputa. Um evento contínuo é aquele que ocorre novamente para o mesmo componente do cluster dentro de um intervalo de cinco minutos e permanece no estado ativo. Para eventos contínuos, o Unified Manager aciona um alerta após detetar o mesmo evento durante dois intervalos de análise consecutivos.

Quando um evento é resolvido, ele permanece disponível no Unified Manager como parte do Registro de problemas de desempenho anteriores de um volume. Cada evento tem um ID exclusivo que identifica o tipo de evento e os volumes, o cluster e os componentes do cluster envolvidos.

Um único volume pode ser envolvido em mais de um evento ao mesmo tempo.

Estado do evento

Os eventos podem estar em um dos seguintes estados:

- **Ativo**

  Indica que o evento de desempenho está ativo no momento (novo ou confirmado). O problema que causa o evento não foi corrigido ou não foi resolvido. O contador de performance do objeto de storage permanece acima do limite de performance.

- **Obsoleto**

  Indica que o evento já não está ativo. O problema que causa o evento foi corrigido ou foi resolvido. O contador de desempenho do objeto de storage não está mais acima do limite de desempenho.

Notificação de evento

Os eventos são exibidos na página Painel e em muitas outras páginas na interface do usuário, e os alertas desses eventos são enviados para endereços de e-mail especificados. Você pode exibir informações de análise detalhadas sobre um evento e obter sugestões para resolvê-lo na página de detalhes do evento e na página análise de carga de trabalho.

Interação de eventos

Na página Detalhes do evento e na página análise de carga de trabalho, você pode interagir com eventos das seguintes maneiras:

- Mover o Mouse sobre um evento exibe uma mensagem que mostra a data e a hora em que o evento foi detetado.

  Se houver vários eventos para o mesmo período de tempo, a mensagem mostrará o número de eventos.

- Clicar em um único evento exibe uma caixa de diálogo que mostra informações mais detalhadas sobre o evento, incluindo os componentes do cluster envolvidos.

  O componente em contenção é circulado e realçado a vermelho. Você pode clicar em **Exibir análise completa** para ver a análise completa na página de detalhes do evento. Se houver vários eventos para o mesmo período de tempo, a caixa de diálogo mostra detalhes sobre os três eventos mais recentes. Você pode clicar em um evento para exibir a análise do evento na página de detalhes do evento.

Como o Unified Manager determina o impacto no desempenho de um evento

O Unified Manager usa o desvio na atividade, utilização, taxa de transferência de gravação, uso de componentes do cluster ou latência de e/S (tempo de resposta) para um workload a fim de determinar o nível de impacto na performance do workload. Essas informações determinam a função de cada carga de trabalho no evento e como eles são classificados na página de detalhes do evento.

O Unified Manager compara os últimos valores analisados para uma carga de trabalho com o intervalo esperado (previsão de latência) de valores. A diferença entre os valores analisados pela última vez e o intervalo esperado de valores identifica as cargas de trabalho cujo desempenho foi mais impactado pelo evento.

Por exemplo, suponha que um cluster contenha duas cargas de trabalho: Carga de Trabalho A e carga de trabalho B. a previsão de latência para carga de trabalho A é de 5-10 milissegundos por operação (ms/op) e sua latência real geralmente é de cerca de 7 ms/op. A previsão de latência para o workload B é de 10-20 ms/op e sua latência real geralmente é de cerca de 15 ms/op. Ambos os workloads estão bem dentro da previsão de latência. Devido à contenção no cluster, a latência de ambas as cargas de trabalho aumenta para 40 ms/op, cruzando o limite de desempenho dinâmico, que é os limites superiores da previsão de latência e acionando eventos. O desvio de latência, dos valores esperados para os valores acima do limite de desempenho, para a carga de trabalho A é de cerca de 33 ms/op, e o desvio para carga de trabalho B é de cerca de 25 ms/op. A latência de ambas as cargas de trabalho aumenta para 40 ms/op, mas o Workload A teve maior impacto no desempenho, pois teve maior desvio de latência em 33 ms/op.

Na página de detalhes do evento, na seção Diagnóstico do sistema, você pode classificar as cargas de trabalho por seu desvio na atividade, utilização ou taxa de transferência de um componente do cluster. Você também pode classificar workloads por latência. Quando você seleciona uma opção de classificação, o Unified Manager analisa o desvio na atividade, utilização, taxa de transferência ou latência desde que o evento foi detetado a partir dos valores esperados para determinar a ordem de classificação da carga de trabalho. Para a latência, os pontos vermelhos (🛑) indicam um cruzamento de limite de desempenho por uma carga de trabalho da vítima e o impactos subsequente na latência. Cada ponto vermelho indica um nível mais alto de desvio na latência, o que ajuda a identificar as cargas de trabalho da vítima cuja latência foi mais afetada por um evento.

Componentes do cluster e por que eles podem estar na contenção

Você pode identificar problemas de desempenho do cluster quando um componente do cluster entra em contenção. O desempenho de workloads que usam o componente diminui e seu tempo de resposta (latência) para solicitações do cliente aumenta, o que aciona um evento no Unified Manager.

Um componente que está em disputa não pode funcionar em um nível ideal. Seu desempenho diminuiu e o desempenho de outros componentes e cargas de trabalho do cluster, chamados *vítimas*, pode ter aumentado a latência. Para sair da contenção de um componente, você precisa reduzir o workload ou aumentar a

capacidade de lidar com mais trabalho, para que a performance possa retornar aos níveis normais. Como o Unified Manager coleta e analisa a performance do workload em intervalos de cinco minutos, ele deteta quando um componente do cluster é consistentemente sobreusado. Não são detetados picos transitórios de sobreutilização que duram apenas uma curta duração dentro do intervalo de cinco minutos.

Por exemplo, um agregado de storage pode estar sob contenção porque um ou mais workloads nele estão competindo para que suas solicitações de e/S sejam atendidas. Outras cargas de trabalho no agregado podem ser afetadas, fazendo com que seu desempenho diminua. Para reduzir a quantidade de atividade no agregado, há etapas diferentes, como mover uma ou mais workloads para um agregado ou nó menos ocupado, para diminuir a demanda geral de workload no agregado atual. Para um grupo de políticas de QoS, você pode ajustar o limite de taxa de transferência ou mover workloads para um grupo de políticas diferente, para que os workloads não fiquem mais sendo controlados.

O Unified Manager monitora os seguintes componentes do cluster para alertá-lo quando eles estão na contenção:

- **Rede**

  Representa o tempo de espera das solicitações de e/S pelos protocolos de rede externos no cluster. O tempo de espera é o tempo gasto esperando que as transações "prontas para transferência" sejam concluídas antes que o cluster possa responder a uma solicitação de e/S. Se o componente de rede estiver em contenção, isso significa que o alto tempo de espera na camada de protocolo está impactando a latência de uma ou mais cargas de trabalho.

- **Processamento de rede**

  Representa o componente de software no cluster envolvido com o processamento de e/S entre a camada de protocolo e o cluster. O processamento da rede de tratamento do nó pode ter sido alterado desde que o evento foi detetado. Se o componente de processamento de rede estiver em contenção, isso significa que a alta utilização no nó de processamento de rede está impactando a latência de uma ou mais cargas de trabalho.

  Ao usar um cluster All SAN Array em uma configuração ativo-ativo, o valor de latência de processamento de rede é exibido para ambos os nós para que você possa verificar se os nós estão compartilhando a carga igualmente.

- * Limite de QoS Max*

  Representa a configuração de taxa de transferência máxima (pico) do grupo de políticas de qualidade do serviço (QoS) de storage atribuído ao workload. Se o componente do grupo de políticas estiver na contenção, isso significa que todas as cargas de trabalho no grupo de políticas estão sendo controladas pelo limite de taxa de transferência definido, o que está impactando a latência de uma ou mais dessas cargas de trabalho.

- **Limite de QoS min**

  Representa a latência de um workload que está sendo causado pela configuração mínima (esperada) de taxa de transferência de QoS atribuída a outros workloads. Se o conjunto mínimo de QoS em certos workloads usar a maior parte da largura de banda para garantir a taxa de transferência prometida, outros workloads serão controlados e verão mais latência.

- **Interconexão de cluster**

  Representa os cabos e adaptadores com os quais os nós em cluster estão fisicamente conetados. Se o componente de interconexão de cluster estiver na contenção, isso significa que o tempo de espera alto

para solicitações de e/S na interconexão de cluster está impactando a latência de um ou mais workloads.

- **Data Processing**

  Representa o componente de software no cluster envolvido com o processamento de e/S entre o cluster e o agregado de storage que contém a carga de trabalho. O Data Processing de tratamento do nó pode ter sido alterado desde que o evento foi detetado. Se o componente Data Processing estiver em contenção, isso significa que a alta utilização no nó Data Processing está impactando a latência de um ou mais workloads.

- * Ativação de volume*

  Representa o processo que controla o uso de todos os volumes ativos. Em ambientes grandes onde mais de 1000 volumes estão ativos, esse processo controla quantos volumes críticos precisam acessar recursos por meio do nó ao mesmo tempo. Quando o número de volumes ativos simultâneos exceder o limite máximo recomendado, alguns dos volumes não críticos terão latência conforme identificado aqui.

- **Recursos MetroCluster**

  Representa os recursos do MetroCluster, incluindo NVRAM e links interswitches (ISLs), usados para espelhar dados entre clusters em uma configuração do MetroCluster. Se o componente MetroCluster estiver em contenção, isso significa que a alta taxa de transferência de gravação de workloads no cluster local ou um problema de integridade de link está impactando a latência de um ou mais workloads no cluster local. Se o cluster não estiver em uma configuração do MetroCluster, este ícone não será exibido.

- **Operações agregadas ou SSD agregadas**

  Representa o agregado de storage no qual os workloads estão sendo executados. Se o componente agregado estiver na contenção, isso significa que a alta utilização no agregado está impactando a latência de um ou mais workloads. Um agregado consiste em todos os HDDs, ou uma combinação de HDDs e SSDs (agregado de Flash Pool), ou uma combinação de HDDs e uma camada de nuvem (agregado de FabricPool). Um "agregado SSD" consiste em todos os SSDs (um agregado all-flash) ou uma combinação de SSDs e uma camada de nuvem (agregado FabricPool).

- **Latência da nuvem**

  Representa o componente de software no cluster envolvido com o processamento de e/S entre o cluster e a camada de nuvem na qual os dados do usuário são armazenados. Se o componente de latência da nuvem estiver em contenção, isso significa que uma grande quantidade de leituras de volumes hospedados na camada de nuvem está impactando a latência de um ou mais workloads.

- **Sincronizar SnapMirror**

  Representa o componente de software no cluster envolvido com a replicação dos dados do usuário do volume primário para o volume secundário em uma relação síncrona do SnapMirror. Se o componente Sync SnapMirror estiver na contenção, isso significa que a atividade das operações síncronas do SnapMirror está impactando a latência de um ou mais workloads.

Funções dos workloads envolvidos em um evento de desempenho

O Unified Manager usa funções para identificar o envolvimento de um workload em um evento de performance. Os papéis incluem vítimas, agressores e tubarões. Uma carga de trabalho definida pelo usuário pode ser uma vítima, um valentão e um tubarão ao mesmo tempo.

Função	Descrição
Vítima	Uma carga de trabalho definida pelo usuário cujo desempenho diminuiu devido a outras cargas de trabalho, chamadas de bullies, que usam excessivamente um componente de cluster. Somente cargas de trabalho definidas pelo usuário são identificadas como vítimas. O Unified Manager identifica os workloads da vítima com base em seu desvio na latência, em que a latência real, durante um evento, aumentou muito em relação à previsão de latência (intervalo esperado).
Bully	Uma carga de trabalho definida pelo usuário ou definida pelo sistema cujo uso excessivo de um componente de cluster causou a diminuição do desempenho de outras cargas de trabalho, chamadas vítimas. O Unified Manager identifica cargas de trabalho bully com base em seu desvio no uso de um componente do cluster, em que o uso real, durante um evento, aumentou muito em relação ao intervalo de uso esperado.
Tubarão	Um workload definido pelo usuário com a maior utilização de um componente de cluster em comparação a todas as cargas de trabalho envolvidas em um evento. O Unified Manager identifica workloads de tubarão com base no uso de um componente de cluster durante um evento.

Os workloads em um cluster podem compartilhar muitos dos componentes do cluster, como agregados e CPU para rede e Data Processing. Quando uma carga de trabalho, como um volume, aumenta o uso de um componente de cluster a ponto de que o componente não pode atender com eficiência às demandas de workload, o componente está em contenção. A carga de trabalho que está usando um componente de cluster é um bully. As outras cargas de trabalho que compartilham esses componentes, e cujo desempenho é afetado pelo agressor, são as vítimas. As atividades de workloads definidos pelo sistema, como deduplicação ou cópias Snapshot, também podem escalar para "bullying".

Quando o Unified Manager deteta um evento, ele identifica todos os workloads e componentes de cluster envolvidos, incluindo os workloads bully que causaram o evento, o componente do cluster que está em contenção e os workloads da vítima cujo desempenho diminuiu devido ao aumento da atividade dos workloads bully.

Se o Unified Manager não conseguir identificar os workloads bully, ele só alertará sobre os workloads da vítima e o componente do cluster envolvido.

O Unified Manager pode identificar workloads vítimas de workloads bully e também identificar quando esses mesmos workloads se tornam workloads bully. Uma carga de trabalho pode ser um bully para si mesma. Por exemplo, uma carga de trabalho de alta performance que está sendo controlada por um limite de grupo de políticas faz com que todas as cargas de trabalho no grupo de políticas sejam limitadas, incluindo a própria. Uma carga de trabalho que é um agressor ou uma vítima em um evento de desempenho contínuo pode mudar sua função ou não ser mais um participante no evento.

Informações sobre direitos autorais

Informações sobre marcas comerciais