Recogida de datos y supervisión del rendimiento de carga de trabajo

Active IQ Unified Manager 9.8

NetApp
April 16, 2024

Tabla de contenidos

Recogida de datos y supervisión del rendimiento de carga de trabajo

Unified Manager recopila y analiza la actividad de la carga de trabajo cada 5 minutos para identificar eventos de rendimiento y detecta cambios de configuración cada 15 minutos. Conserva un máximo de 30 días de datos históricos de eventos y rendimiento de 5 minutos, y utiliza estos datos para prever el rango de latencia esperado para todas las cargas de trabajo supervisadas.

Unified Manager debe recoger como mínimo 3 días de actividad de carga de trabajo antes de que pueda iniciar su análisis y antes de que la previsión de latencia del tiempo de respuesta de I/o pueda mostrarse en la página Workload Analysis y en la página de detalles Event. Mientras se recoge esta actividad, la previsión de latencia no muestra todos los cambios que se producen en la actividad de la carga de trabajo. Después de recoger 3 días de actividad, Unified Manager ajusta el pronóstico de latencia cada 24 horas a las 12:00 a.m. para reflejar los cambios en la actividad de la carga de trabajo y establecer un umbral dinámico de rendimiento más preciso.

Durante los primeros 4 días en los que Unified Manager supervisa una carga de trabajo, si han transcurrido más de 24 horas desde la última recogida de datos, los gráficos de latencia no mostrarán el pronóstico de latencia de esa carga de trabajo. Los eventos detectados antes de la última recogida siguen estando disponibles.

> El horario de verano cambia la hora del sistema, lo que altera el pronóstico de latencia de las estadísticas de rendimiento para las cargas de trabajo supervisadas. Unified Manager empieza a corregir el pronóstico de latencia de forma inmediata y tarda unos 15 días en completarlo. Durante este tiempo, puede seguir usando Unified Manager, pero, dado que Unified Manager utiliza el pronóstico de latencia para detectar eventos dinámicos, es posible que algunos eventos no sean precisos. Los eventos detectados antes del cambio de hora no se verán afectados.

Tipos de cargas de trabajo supervisadas por Unified Manager

Es posible utilizar Unified Manager para supervisar el rendimiento de dos tipos de cargas de trabajo: Definidas por el usuario y definidas por el sistema.

- ***cargas de trabajo definidas por el usuario***

  El rendimiento de I/o de las aplicaciones al clúster. Estos son procesos implicados en solicitudes de lectura y escritura. Un volumen, LUN, recurso compartido NFS, recurso compartido SMB/CIFS y una carga de trabajo es una carga de trabajo definida por el usuario.

  > Unified Manager solo supervisa la actividad de las cargas de trabajo en el clúster. No supervisa las aplicaciones, los clientes ni las rutas entre las aplicaciones y el clúster.

  Si se cumple uno o varios de los siguientes aspectos con relación a una carga de trabajo, Unified Manager no se podrá supervisar:

  - Se trata de una copia de protección de datos (DP) en modo de solo lectura. (Los volúmenes DP se

supervisan para el tráfico generado por el usuario).

- Se trata de un clon de datos sin conexión.
- Es un volumen reflejado en una configuración MetroCluster.

- ***cargas de trabajo definidas por el sistema***

  Los procesos internos implicados en la eficiencia del almacenamiento, la replicación de datos y el estado del sistema, incluidos:

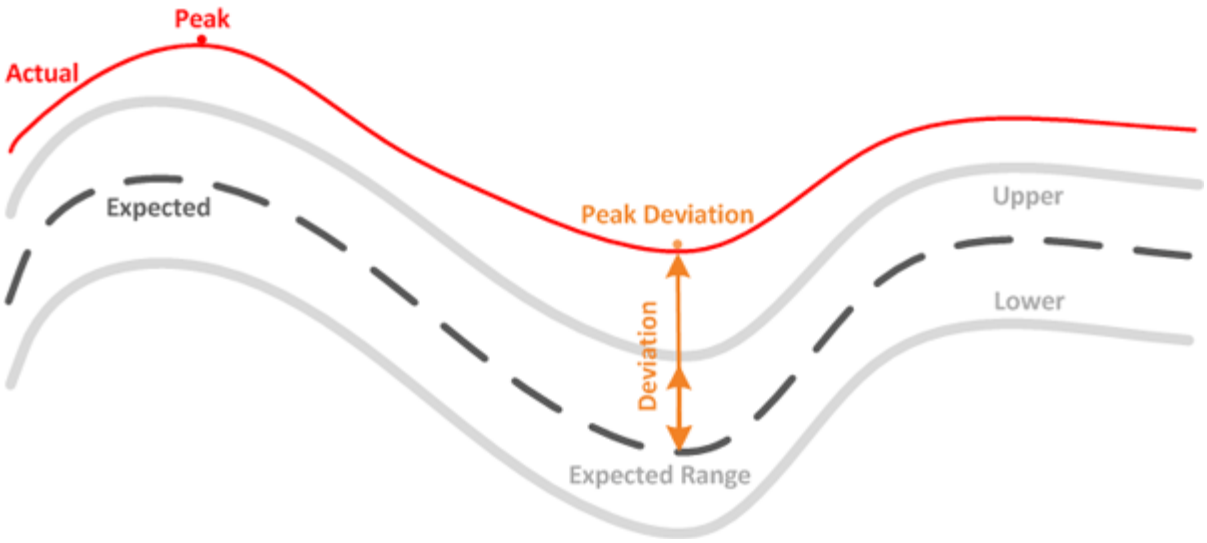
  - Eficiencia del almacenamiento, como la deduplicación
  - Estado del disco, que incluye la reconstrucción de RAID, el filtrado de discos, etc.
  - La replicación de datos, como las copias SnapMirror
  - Actividades de gestión
  - Estado del sistema de archivos, que incluye diversas actividades de WAFL
  - Escáneres del sistema de archivos, como WAFL scan
  - Descarga de copias, como las operaciones de eficiencia del almacenamiento descargados de hosts VMware
  - Estado del sistema, como movimientos de volúmenes, compresión de datos, etc.
  - Volúmenes sin supervisar

Los datos de rendimiento de las cargas de trabajo definidas por el sistema se muestran en la interfaz gráfica de usuario solo cuando el componente del clúster utilizado por estas cargas de trabajo es objeto de disputa. Por ejemplo, no se puede buscar el nombre de una carga de trabajo definida por el sistema para ver sus datos de rendimiento en la interfaz gráfica de usuario.

Valores de medición del rendimiento de la carga de trabajo

Unified Manager mide el rendimiento de las cargas de trabajo en un clúster en función de los valores estadísticos históricos y esperados, lo que forma el pronóstico de latencia de los valores de las cargas de trabajo. Compara los valores estadísticos de la carga de trabajo real con el pronóstico de latencia para determinar cuándo el rendimiento de la carga de trabajo es demasiado alto o demasiado bajo. Una carga de trabajo que no se ejecuta según lo esperado activa un evento de rendimiento dinámico para notificarle.

En la siguiente ilustración, el valor real, en rojo, representa las estadísticas de rendimiento reales en el plazo. El valor real ha superado el umbral de rendimiento, que es los límites superiores del pronóstico de latencia. El pico es el valor real más alto del lapso. La desviación mide el cambio entre los valores esperados (el pronóstico) y los valores reales, mientras que la desviación máxima indica el mayor cambio entre los valores esperados y los valores reales.



En la siguiente tabla, se enumeran los valores de medición del rendimiento de la carga de trabajo.

Medida	Descripción
Actividad	El porcentaje del límite de calidad de servicio que utilizan las cargas de trabajo del grupo de políticas. Si Unified Manager detecta un cambio en un grupo de políticas, como añadir o quitar un volumen o cambiar el límite de calidad de servicio, los valores reales y esperados pueden superar el 100 % del límite establecido. Si un valor supera el 100% del límite establecido, se muestra como >100%. Si un valor es inferior al 1% del límite establecido, se muestra como <1%.
Real	El valor de rendimiento medido en un momento específico para una carga de trabajo determinada.
Desviación	El cambio entre los valores esperados y los valores reales. Es la relación entre el valor real menos el valor esperado y el valor superior del rango esperado menos el valor esperado. Un valor de desviación negativo indica que el rendimiento de la carga de trabajo es menor de lo esperado, mientras que un valor positivo indica que el rendimiento de la carga de trabajo es mayor de lo esperado.

Medida	Descripción
Esperado	Los valores esperados se basan en el análisis de los datos de rendimiento históricos de una carga de trabajo determinada. Unified Manager analiza estos valores estadísticos para determinar el rango esperado (previsión de latencia) de valores.
Previsión de latencia (rango esperado)	La previsión de latencia es una predicción de los valores de rendimiento superior e inferior esperados en un momento determinado. Para la latencia de la carga de trabajo, los valores superiores forman el umbral de rendimiento. Cuando el valor real supera el umbral de rendimiento, Unified Manager activa un evento de rendimiento dinámico.
Pico	Valor máximo medido a lo largo de un período de tiempo.
Desviación máxima	El valor máximo de desviación medido a lo largo de un período de tiempo.
Profundidad de cola	El número de solicitudes de I/o pendientes que están esperando en el componente de interconexión.
Utilización	Para los componentes de procesamiento de red, procesamiento de datos y agregado, el porcentaje de tiempo ocupado para completar las operaciones de carga de trabajo durante un período de tiempo. Por ejemplo, el porcentaje de tiempo durante el procesamiento de la red o los componentes de procesamiento de datos para procesar una solicitud de I/o o o para que un agregado cumpla una solicitud de lectura o escritura.
Rendimiento de escritura	La cantidad de rendimiento de escritura, en megabytes por segundo (MB/s), desde cargas de trabajo en un clúster local hasta el clúster asociado en una configuración MetroCluster.

Qué rango esperado de rendimiento es

La previsión de latencia es una predicción de los valores de rendimiento superior e inferior esperados en un momento determinado. Para la latencia de la carga de trabajo, los valores superiores forman el umbral de rendimiento. Cuando el valor real supera el umbral de rendimiento, Unified Manager activa un evento de rendimiento dinámico.

Por ejemplo, durante el horario normal de trabajo entre las 9:00 a.m. y a las 5:00 p.m., la mayoría de los empleados podrían revisar su correo electrónico entre las 9:00 a.m. y las 10:30 El aumento de la demanda de

los servidores de correo electrónico implica un aumento de la actividad de las cargas de trabajo en el almacenamiento de entorno de administración durante este tiempo. Los empleados pueden notar que el tiempo de respuesta es lento por parte de los clientes de correo electrónico.

Durante la hora del almuerzo entre las 12:00 p.m. y la 1:00 y al final del día de trabajo después de las 5:00 p.m., la mayoría de los empleados están probablemente lejos de sus computadoras. La demanda de los servidores de correo electrónico suele disminuir y, por lo tanto, se reduce la demanda de almacenamiento de entorno administrativo. Como alternativa, podrían programarse operaciones con la carga de trabajo, como backups de almacenamiento o análisis de virus, que comiencen a las 5:00 p. m. y aumente la actividad en el almacenamiento de gestión.

Durante varios días, el aumento y la disminución de la actividad de las cargas de trabajo determinan el rango esperado (pronóstico de latencia) de la actividad, con los límites superior e inferior de una carga de trabajo. Cuando la actividad de carga de trabajo real de un objeto está fuera de los límites superior o inferior, y permanece fuera de los límites durante un periodo de tiempo, puede indicar que el objeto se está utilizando en exceso o infrautilizado.

Cómo se forma la previsión de latencia

Unified Manager debe recoger como mínimo 3 días de actividad de carga de trabajo antes de que pueda iniciar su análisis y antes de que la previsión de latencia del tiempo de respuesta de I/o pueda mostrarse en la interfaz gráfica de usuario. La recopilación de datos mínimos requeridos no tiene en cuenta todos los cambios que se producen en la actividad de la carga de trabajo. Después de recoger los primeros 3 días de actividad, Unified Manager ajusta el pronóstico de latencia cada 24 horas a las 12:00 para reflejar los cambios en la actividad de las cargas de trabajo y establecer un umbral dinámico de rendimiento más preciso.

> El horario de verano cambia la hora del sistema, lo que altera el pronóstico de latencia de las estadísticas de rendimiento para las cargas de trabajo supervisadas. Unified Manager empieza a corregir el pronóstico de latencia de forma inmediata y tarda unos 15 días en completarlo. Durante este tiempo, puede seguir usando Unified Manager, pero, dado que Unified Manager utiliza el pronóstico de latencia para detectar eventos dinámicos, es posible que algunos eventos no sean precisos. Los eventos detectados antes del cambio de hora no se verán afectados.

Cómo se utiliza el pronóstico de latencia en el análisis de rendimiento

Unified Manager usa el pronóstico de latencia para representar la actividad de latencia de I/o típica (tiempo de respuesta) para las cargas de trabajo supervisadas. Le avisa cuando la latencia real de una carga de trabajo se encuentra por encima de los límites superiores de la previsión de latencia, que activa un evento de rendimiento dinámico, para que pueda analizar el problema de rendimiento y tomar medidas correctivas para resolverlo.

El pronóstico de latencia establece la línea de base de rendimiento para la carga de trabajo. Con el tiempo, Unified Manager aprende de las mediciones de rendimiento anteriores para pronosticar los niveles de rendimiento y actividad esperados para la carga de trabajo. El límite superior del rango esperado establece el umbral de rendimiento dinámico. Unified Manager utiliza la base para determinar cuándo la latencia real está por encima o por debajo de un umbral, o fuera de los límites de su rango esperado. La comparación entre los valores reales y los valores esperados crea un perfil de rendimiento para la carga de trabajo.

Cuando la latencia real de una carga de trabajo supera el umbral de rendimiento dinámico, debido a la contención en un componente de clúster, la latencia es alta y la carga de trabajo se ejecuta más lentamente de lo esperado. El rendimiento de otras cargas de trabajo que comparten los mismos componentes del clúster también puede ser más lento de lo esperado.

Unified Manager analiza el evento de umbral que cruza y determina si la actividad es un evento de rendimiento. Si la actividad de alta carga de trabajo continúa siendo coherente durante un largo periodo de tiempo, como varias horas, Unified Manager considera que la actividad es normal y ajusta de forma dinámica el pronóstico de latencia para formar el nuevo umbral de rendimiento dinámico.

Es posible que algunas cargas de trabajo tengan una actividad baja constante, donde la previsión de latencia no tiene una tasa alta de cambio con el tiempo. Para minimizar el número de eventos durante el análisis de eventos de rendimiento, Unified Manager activa un evento solo en volúmenes de baja actividad cuyas operaciones y latencias son mucho más elevadas de lo esperado.

En este ejemplo, la latencia de un volumen tiene un pronóstico de latencia, en gris, de 3.5 milisegundos por operación (ms/op) a su nivel más bajo y de 5.5 ms/op a su nivel más alto. Si la latencia real, en azul, aumenta de repente a 10 ms/op, debido a un pico intermitente en el tráfico de la red o contención en un componente del clúster, supera la previsión de latencia y ha superado el umbral de rendimiento dinámico.

Cuando se redujo el tráfico de red o el componente del clúster ya no es objeto de disputa, la latencia devuelve dentro de la previsión de latencia. Si la latencia permanece en o por encima de los 10 ms/op durante un largo periodo de tiempo, es posible que deba tomar la acción correctiva para resolver el evento.

Cómo Unified Manager utiliza la latencia de carga de trabajo para identificar problemas de rendimiento

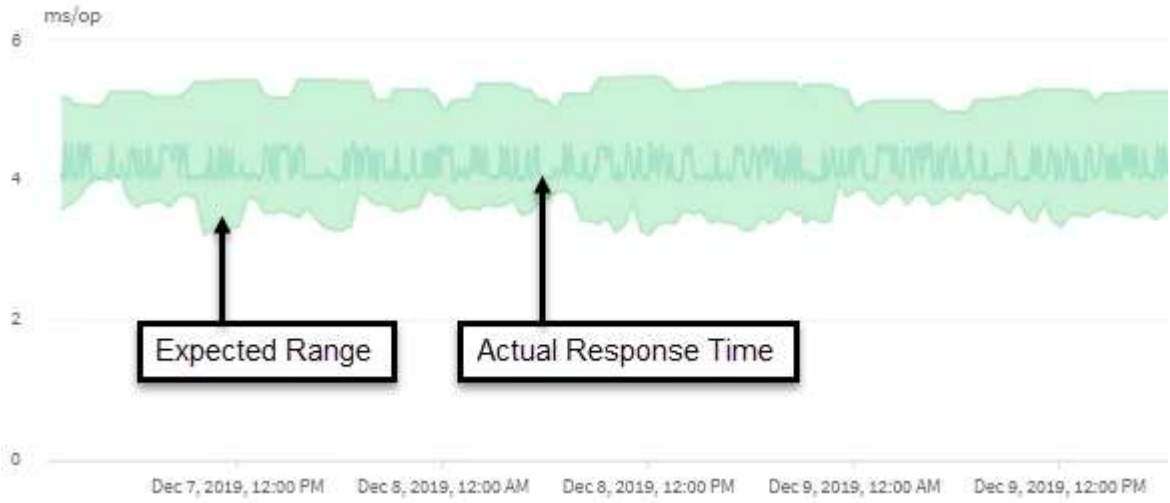
La latencia de las cargas de trabajo (tiempo de respuesta) es el tiempo que tarda un volumen en un clúster en responder a las solicitudes de I/o de las aplicaciones cliente. Unified Manager utiliza la latencia para detectar eventos de rendimiento y alertarle de ellos.

Una alta latencia significa que las solicitudes de aplicaciones a un volumen de un clúster tardan más de lo habitual. La causa de la alta latencia podría estar en el clúster mismo, debido a la contención en uno o más componentes del clúster. La alta latencia también podría deberse a problemas externos al clúster, como cuellos de botella de red, problemas con el cliente que aloja las aplicaciones, o problemas con las propias aplicaciones.

Unified Manager solo supervisa la actividad de las cargas de trabajo en el clúster. No supervisa las aplicaciones, los clientes ni las rutas entre las aplicaciones y el clúster.

Las operaciones en el clúster, como la realización de backups o la ejecución de la deduplicación, que aumenten su demanda de los componentes del clúster compartidos por otras cargas de trabajo, también pueden contribuir a una alta latencia. Si la latencia real supera el umbral de rendimiento dinámico del rango esperado (previsión de latencia), Unified Manager analiza el evento para determinar si es un evento de rendimiento que podría necesitar resolver. La latencia se mide en milisegundos por operación (ms/op).

En el gráfico Latency total de la página Workload Analysis, es posible ver un análisis de las estadísticas de latencia para ver cómo la actividad de procesos individuales, como solicitudes de lectura y escritura, se compara con las estadísticas de latencia generales. La comparación ayuda a determinar qué operaciones tienen la actividad más alta o si operaciones específicas tienen actividad anormal que afecta la latencia de un volumen. Al analizar los eventos de rendimiento, puede utilizar las estadísticas de latencia para determinar si un evento fue causado por un problema en el clúster. También puede identificar las actividades de carga de trabajo específicas o los componentes del clúster que participan en el evento.



En este ejemplo se muestra el gráfico latencia . La actividad de tiempo de respuesta real (latencia) es una línea azul y la previsión de latencia (rango esperado) es verde.

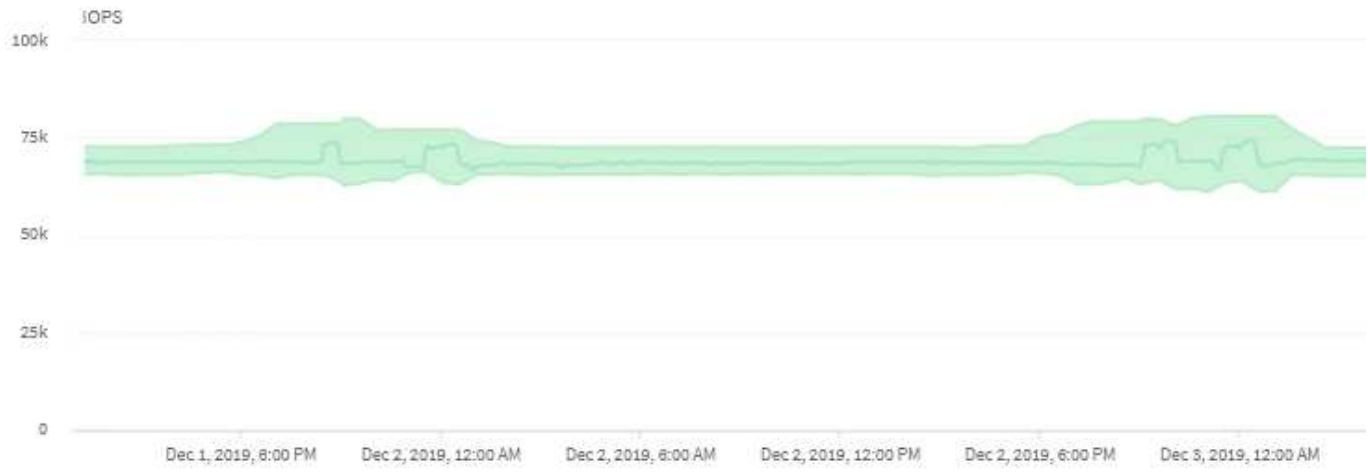
Puede haber vacíos en la línea azul si Unified Manager no pudo recopilar datos. Esto puede ocurrir porque el clúster o el volumen no se pudo acceder, Unified Manager se apagó durante esa hora o la recogida fue más prolongada que el período de recogida de 5 minutos.

Cómo las operaciones del clúster pueden afectar a la latencia de carga de trabajo

Las operaciones (IOPS) representan la actividad de todas las cargas de trabajo definidas por el usuario y definidas por el sistema en un clúster. Las estadísticas de IOPS le ayudan a determinar si los procesos del clúster, como hacer backups o ejecutar deduplicación, están afectando a la latencia de las cargas de trabajo (tiempo de respuesta) o podrían haber provocado o contribuido a un evento de rendimiento.

Al analizar los eventos de rendimiento, puede utilizar las estadísticas de IOPS para determinar si un evento de rendimiento fue causado por un problema en el clúster. Es posible identificar las actividades de carga de

trabajo específicas que pueden haber contribuido principalmente al evento de rendimiento. Las IOPS se miden en operaciones por segundo (OPS/s).



Este ejemplo muestra el gráfico IOPS. Las estadísticas de operaciones reales son una línea azul y las estadísticas de operaciones de IOPS se muestran en verde.

En algunos casos en los que se sobrecarga un clúster, es posible que Unified Manager muestre el mensaje `Data collection is taking too long on Cluster cluster_name`. Esto significa que no se han recopilado suficientes estadísticas para que Unified Manager las analice. Debe reducir los recursos que está utilizando el clúster para poder recopilar las estadísticas.

Supervisión del rendimiento de las configuraciones de MetroCluster

Unified Manager permite supervisar el rendimiento de escritura entre clústeres en una configuración de MetroCluster para identificar cargas de trabajo con un alto nivel de rendimiento de escritura. Si estas cargas de trabajo de alto rendimiento hacen que otros volúmenes en el clúster local tengan tiempos de respuesta de I/o elevados, Unified Manager activa los eventos de rendimiento para notificarle.

Cuando un clúster local en una configuración MetroCluster refleja sus datos en el clúster del partner, los datos se escriben en NVRAM y, después, se transfieren mediante los vínculos interswitch (ISL) a los agregados remotos. Unified Manager analiza la NVRAM para identificar las cargas de trabajo cuyo alto rendimiento de escritura está utilizando demasiado la NVRAM, poniendo en contención a la NVRAM.

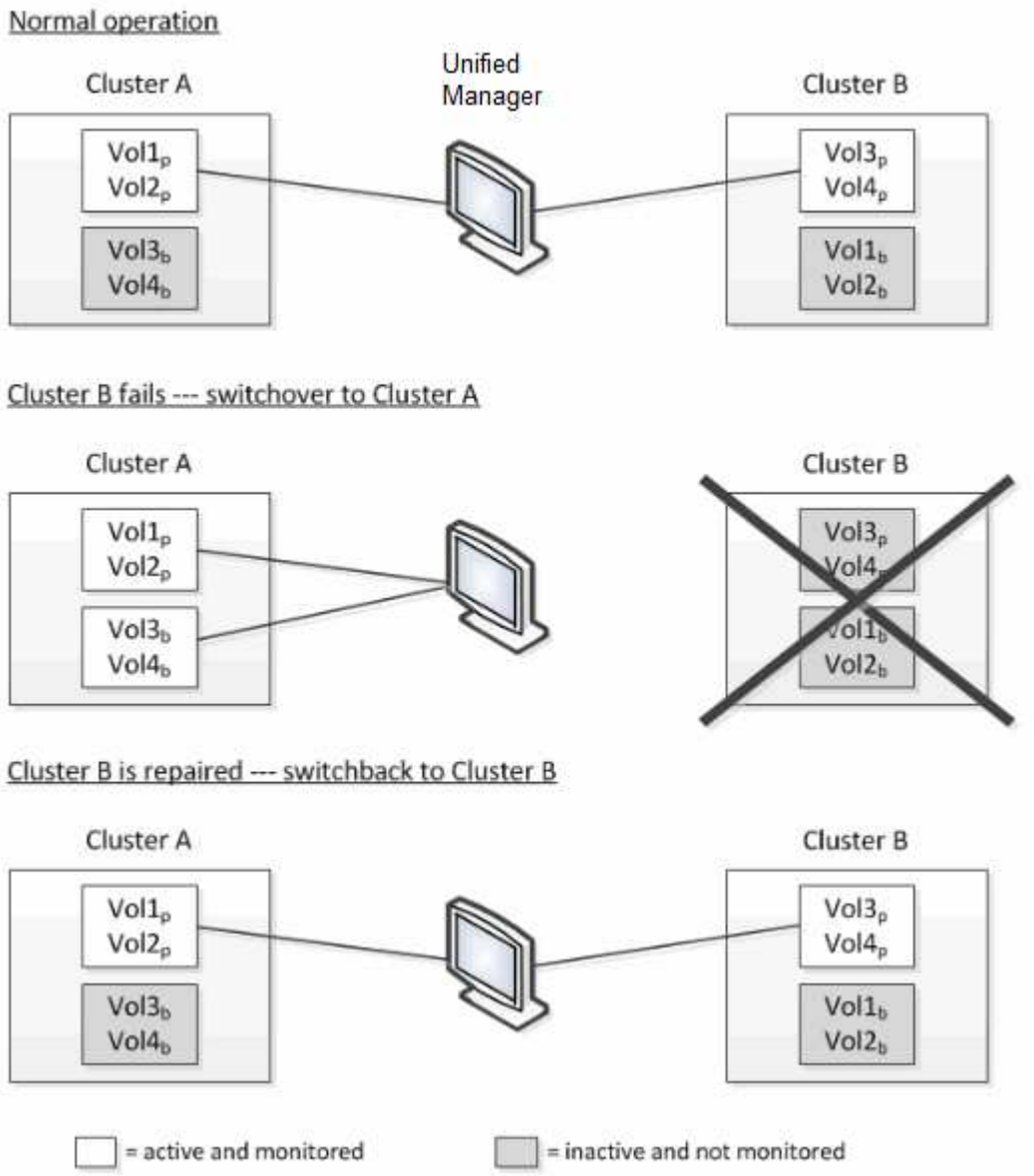
Las cargas de trabajo cuya desviación en el tiempo de respuesta ha superado el umbral de rendimiento se denominan *víctimas* y las cargas de trabajo cuya desviación en el rendimiento de escritura en la NVRAM es mayor de lo habitual, lo que provoca la contención, se denominan *abusones*. Dado que solo se reflejan las solicitudes de escritura en el clúster de partners, Unified Manager no analiza el rendimiento de lectura.

Unified Manager trata los clústeres de una configuración de MetroCluster como clústeres individuales. No distingue entre clústeres que son partners ni relaciona el rendimiento de escritura de cada clúster.

Comportamiento del volumen durante la conmutación de sitios y la conmutación de estado

Los eventos que activan una conmutación de sitios o una conmutación de estado hacen que los volúmenes activos se muevan de un clúster a otro en el grupo de recuperación ante desastres. Se detienen los volúmenes del clúster activos y sirven datos a los clientes, y los volúmenes del otro clúster se activan y comienzan a servir datos. Unified Manager supervisa solo los volúmenes que están activos y en ejecución.

Dado que los volúmenes se mueven de un clúster a otro, se recomienda supervisar ambos clústeres. Una única instancia de Unified Manager puede supervisar ambos clústeres en una configuración de MetroCluster, pero en ocasiones la distancia entre las dos ubicaciones requiere el uso de dos instancias de Unified Manager para supervisar ambos clústeres. En la siguiente figura, se muestra una sola instancia de Unified Manager:



Los volúmenes con p en sus nombres indican los volúmenes primarios, y los volúmenes con b en sus nombres son volúmenes de backup reflejados creados por SnapMirror.

Durante el funcionamiento normal:

- El clúster A tiene dos volúmenes activos: Vol1p y Vol2p.
- El clúster B tiene dos volúmenes activos: Vol3p y Vol4p.
- El clúster A tiene dos volúmenes inactivos: Vol3b y Vol4b.
- El clúster B tiene dos volúmenes inactivos: Vol1b y Vol2b.

Unified Manager recopila la información relativa a cada uno de los volúmenes activos (estadísticas, eventos, etc.). Las estadísticas de Vol1p y Vol2p se recogen en el Grupo A, y las estadísticas de Vol3p y Vol4p se recogen en el Grupo B.

Una vez que se produce un fallo catastrófico, se lleva a cabo una conmutación de los volúmenes activos del clúster B al clúster A:

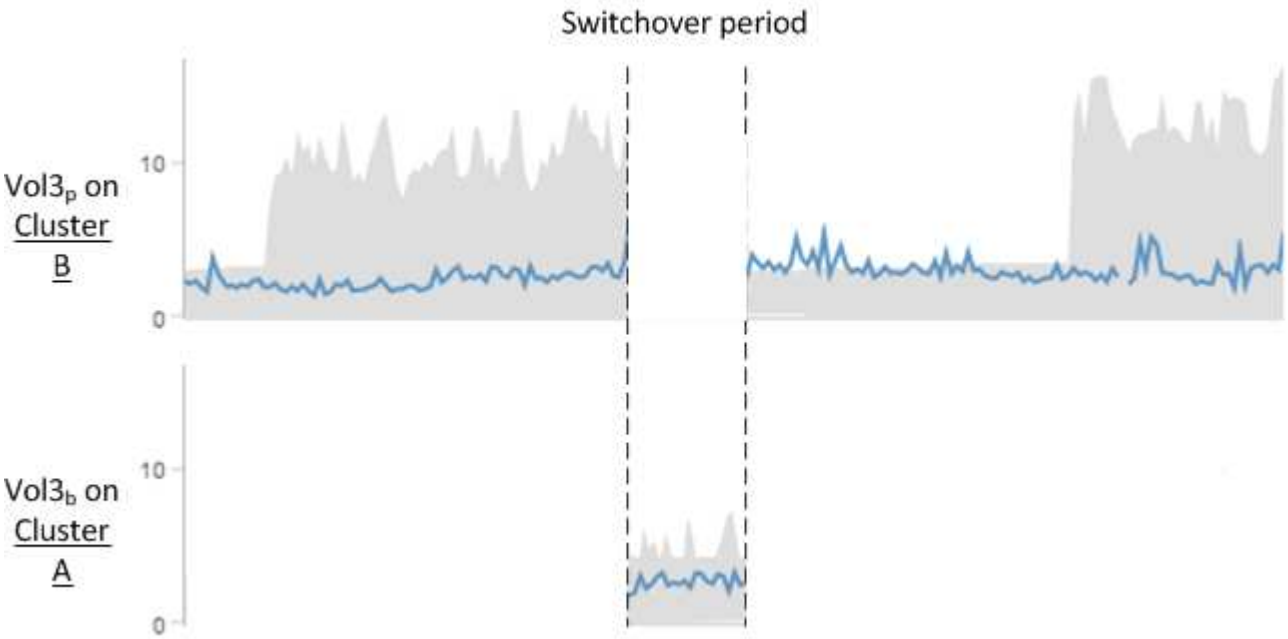
- El clúster A tiene cuatro volúmenes activos: Vol1p, Vol2p, Vol3b y Vol4b.
- El clúster B tiene cuatro volúmenes inactivos: Vol3p, Vol4p, Vol1b y Vol2b.

Como durante el funcionamiento normal, Unified Manager recopila la información relacionada con cada uno de los volúmenes activos. Pero en este caso, las estadísticas de Vol1p y Vol2p son recogidas por el Grupo A, y las estadísticas de Vol3b y Vol4b también son recogidas por el Grupo A.

Tenga en cuenta que Vol3p y Vol3b no son los mismos volúmenes, ya que se encuentran en clústeres diferentes. La información de Unified Manager para Vol3p no es la misma que Vol3b:

- Durante la conmutación al clúster A, las estadísticas y los eventos de Vol3p no se ven.
- En el primer cambio, Vol3b parece un nuevo volumen sin información histórica.

Cuando se repara el clúster B y se realiza una conmutación de regreso, Vol3p vuelve a estar activo en el clúster B, con las estadísticas históricas y una brecha de estadísticas para el período durante el cambio. Vol3b no se puede ver desde el clúster A hasta que se produzca otra conmutación:



- Los volúmenes MetroCluster inactivos, por ejemplo, Vol3b en el clúster A después de la conmutación de estado, se identifican con el mensaje "'este volumen se eliminó'". El volumen no se elimina en realidad, pero Unified Manager no supervisa actualmente porque no es el volumen activo.
- Si una sola instancia de Unified Manager supervisa ambos clústeres en una configuración de MetroCluster, la búsqueda de volúmenes devuelve información de la que esté activo en ese momento. Por ejemplo, una búsqueda para "'Vol3'" devolvería estadísticas y eventos para Vol3b en el clúster A si se ha producido una conmutación y Vol3 se ha vuelto activo en el clúster A.

Análisis y notificación de eventos de rendimiento

Los eventos de rendimiento le notifican problemas de rendimiento de I/o en una carga de trabajo provocada por la contención en un componente de clúster. Unified Manager analiza el evento para identificar todas las cargas de trabajo involucradas, el componente en disputa y si el evento sigue siendo un problema que podría necesitar resolver.

Unified Manager supervisa la latencia de I/o (tiempo de respuesta) y las IOPS (operaciones) para volúmenes en un clúster. Cuando otras cargas de trabajo realizan un uso excesivo de un componente del clúster, por ejemplo, el componente es objeto de disputa y no puede funcionar en un nivel óptimo para satisfacer las demandas de las cargas de trabajo. El rendimiento de otras cargas de trabajo que utilizan el mismo componente puede verse afectado, lo que provoca el aumento de las latencias. Si la latencia supera el umbral de rendimiento dinámico, Unified Manager activa un evento de rendimiento para notificarle.

Análisis de eventos

Unified Manager realiza los siguientes análisis utilizando los 15 días anteriores de estadísticas de rendimiento para identificar las cargas de trabajo víctimas, las cargas de trabajo abusivas y el componente del clúster implicados en un evento:

- Identifica las cargas de trabajo víctimas cuya latencia ha superado el umbral de rendimiento dinámico, que es el límite superior de la previsión de latencia:
  - Para volúmenes en agregados híbridos HDD o Flash Pool (nivel local), los eventos solo se activan cuando la latencia es superior a 5 milisegundos (ms) y las IOPS son superiores a 10 operaciones por segundo (OPS/s).
  - Para volúmenes en agregados íntegramente de SSD o agregados de FabricPool (nivel de cloud), los eventos solo se activan cuando la latencia es superior a 1 ms y el IOPS supera las 100 OPS/s.
- Identifica el componente del clúster en disputa.

  Si la latencia de las cargas de trabajo víctimas en la interconexión del clúster es superior a 1 ms, Unified Manager lo considera importante y activa un evento para la interconexión del clúster.

- Identifica las cargas de trabajo abusivas que están sobreutilizando el componente del clúster y que hacen que estén en contención.
- Clasifica las cargas de trabajo involucradas, en función de su desviación en utilización o actividad de un componente del clúster, para determinar qué elementos agresores tienen el mayor cambio de uso del componente del clúster y qué víctimas son las más afectadas.

Es posible que se produzca un evento solo durante un breve momento y, a continuación, corregirlo después

de que el componente que está utilizando ya no sea objeto de disputa. Un evento continuo es un evento que se vuelve a producir para el mismo componente del clúster en un intervalo de cinco minutos y permanece en el estado activo. En el caso de eventos continuos, Unified Manager activa una alerta tras detectar el mismo evento en dos intervalos de análisis consecutivos.

Cuando se resuelve un evento, este sigue disponible en Unified Manager como parte del registro de problemas de rendimiento anteriores de un volumen. Cada evento tiene un ID único que identifica el tipo de evento y los componentes de volúmenes, clúster y clúster implicados.

Un único volumen puede participar en más de un evento a la vez.

Estado del evento

Los eventos pueden estar en uno de los siguientes estados:

- **Activo**

  Indica que el evento de rendimiento está activo (nuevo o reconocido). El problema que causa el evento no se ha corregido solo o no se ha resuelto. El contador de rendimiento del objeto de almacenamiento sigue por encima del umbral de rendimiento.

- **Obsoleto**

  Indica que el evento no está activo. El problema que causa el evento se ha corregido solo o se ha resuelto. El contador de rendimiento del objeto de almacenamiento ya no está por encima del umbral de rendimiento.

Notificación de eventos

Los eventos se muestran en la página Dashboard y en muchas otras páginas de la interfaz de usuario, y las alertas de esos eventos se envían a direcciones de correo electrónico especificadas. Puede ver información detallada sobre un evento y obtener sugerencias para resolverlo en la página de detalles Event y en la página Workload Analysis.

Interacción de eventos

En la página de detalles Event y en la página Workload Analysis, puede interactuar con los eventos de las siguientes maneras:

- Al mover el ratón sobre un evento se muestra un mensaje que muestra la fecha y la hora en que se detectó el evento.

  Si hay varios eventos para el mismo período de tiempo, el mensaje muestra el número de eventos.

- Al hacer clic en un solo evento se muestra un cuadro de diálogo que muestra información más detallada sobre el evento, incluidos los componentes del clúster implicados.

  El componente objeto de la contención está en un círculo y se resalta en rojo. Puede hacer clic en **Ver análisis completo** para ver el análisis completo en la página de detalles del evento. Si hay varios eventos para el mismo período de tiempo, el cuadro de diálogo muestra detalles acerca de los tres eventos más recientes. Puede hacer clic en un evento para ver el análisis de eventos en la página de detalles Event.

La forma en que Unified Manager determina el impacto en el rendimiento de un evento

Unified Manager utiliza la desviación de la actividad, la utilización, el rendimiento de escritura, el uso de componentes del clúster o la latencia de I/o (tiempo de respuesta) en una carga de trabajo para determinar el nivel de impacto en el rendimiento de la carga de trabajo. Esta información determina el rol de cada carga de trabajo del evento y su clasificación en la página de detalles Event.

Unified Manager compara los valores del último análisis de una carga de trabajo con el rango esperado (pronóstico de latencia) de valores. La diferencia entre los valores que se analizaron por última vez y el rango esperado de valores identifica las cargas de trabajo cuyo rendimiento tuvo un mayor impacto en el evento.

Por ejemplo, supongamos que un clúster contiene dos cargas de trabajo: La carga de trabajo A y la carga de trabajo B. El pronóstico de latencia de la carga de trabajo A es de 5-10 milisegundos por operación (ms/op) y su latencia real suele ser de aproximadamente 7 ms/op. El pronóstico de latencia para la carga de trabajo B es de 10-20 ms/op y su latencia real suele ser de aproximadamente 15 ms/op. Ambas cargas de trabajo están dentro de sus previsiones de latencia. Debido a una contención en el clúster, la latencia de ambas cargas de trabajo aumenta a 40 ms/op, que cruza el umbral de rendimiento dinámico, que es los límites superiores del pronóstico de latencia y se activan eventos. La desviación en latencia, de los valores esperados a los valores por encima del umbral de rendimiento, para la carga De trabajo A es aproximadamente de 33 ms/op y la desviación de la carga de trabajo B es de unos 25 ms/op. La latencia de ambas cargas de trabajo se dispara hasta los 40 ms/op, pero la carga De trabajo A tuvo un impacto en el rendimiento mayor porque tuvo la mayor desviación de latencia en 33 ms/op.

En la página de detalles Event, en la sección System Diagnosis, se pueden ordenar las cargas de trabajo por su desviación de la actividad, la utilización o el rendimiento de un componente del clúster. También es posible ordenar las cargas de trabajo por latencia. Cuando se selecciona una opción de ordenación, Unified Manager analiza la desviación de la actividad, la utilización, el rendimiento o la latencia desde que se detectó el evento de los valores esperados para determinar el orden de clasificación de la carga de trabajo. Para la latencia, los puntos rojos (🔴) indica un umbral de rendimiento que cruza una carga de trabajo víctima y el impacto posterior en la latencia. Cada punto rojo indica un nivel más alto de desviación en latencia, lo que ayuda a identificar las cargas de trabajo víctimas cuya latencia más se vio afectada por un evento.

Los componentes del clúster y el motivo por los que pueden estar en contención

Puede identificar los problemas de rendimiento del clúster cuando un componente del clúster entra en disputa. El rendimiento de las cargas de trabajo que utilizan el componente disminuye y aumenta su tiempo de respuesta (latencia) para las solicitudes de clientes, lo que activa un evento en Unified Manager.

Un componente que está en disputa no puede ejecutarse en un nivel óptimo. Su rendimiento ha disminuido, y el rendimiento de otros componentes del clúster y las cargas de trabajo, llamado *Victimas*, puede haber aumentado la latencia. Para eliminar un componente de la contención, debe reducir su carga de trabajo o aumentar su capacidad para gestionar más trabajo, de modo que el rendimiento pueda volver a los niveles normales. Dado que Unified Manager recopila y analiza el rendimiento de la carga de trabajo en intervalos de cinco minutos, solo detecta cuando un componente de clúster se utiliza de forma coherente en exceso. No se detectan picos transitorios de sobreutilización que duran solo una corta duración dentro del intervalo de cinco minutos.

Por ejemplo, un agregado de almacenamiento puede ser objeto de disputa porque una o más cargas de trabajo en él están compitiendo por sus solicitudes de I/O. Otras cargas de trabajo del agregado pueden verse

afectadas, lo que provoca una disminución del rendimiento. Para reducir la cantidad de actividad del agregado, hay diferentes pasos que se pueden realizar, como mover una o varias cargas de trabajo a agregados o nodos menos ocupados para reducir la demanda general de la carga de trabajo en el agregado actual. En el caso de un grupo de políticas de calidad de servicio, puede ajustar el límite de rendimiento o mover cargas de trabajo a otro grupo de políticas, de modo que las cargas de trabajo ya no se aceleren.

Unified Manager supervisa los siguientes componentes del clúster para alertarle cuando los hay en disputa:

- **Red**

  Representa el tiempo de espera de las solicitudes de I/o de los protocolos de red externos del clúster. El tiempo de espera transcurrido para que las transacciones «'están listas para la transferencia» finalicen antes de que el clúster pueda responder a una solicitud de I/O. Si el componente de red es objeto de disputa, significa que un tiempo de espera elevado en la capa de protocolo está afectando a la latencia de una o más cargas de trabajo.

- **Procesamiento de red**

  Representa el componente de software del clúster involucrado en el procesamiento de I/o entre la capa de protocolo y el clúster. Es posible que el nodo que gestiona el procesamiento de red haya cambiado desde que se detectó el evento. Si el componente de red es objeto de disputa, significa que un uso elevado en el nodo de procesamiento de red está afectando a la latencia de una o más cargas de trabajo.

  Cuando se utiliza un clúster de cabina All SAN en una configuración activo-activo, el valor de latencia de procesamiento de red se muestra en ambos nodos para poder verificar que los nodos comparten la carga igualmente.

- **Límite de QoS máx.**

  Representa la configuración máxima de rendimiento (pico) del grupo de políticas de calidad de servicio del almacenamiento asignado a la carga de trabajo. Si el componente del grupo de políticas es objeto de disputa, significa que el límite de rendimiento establecido está acelerando todas las cargas de trabajo del grupo de políticas, lo que afecta a la latencia de una o más cargas de trabajo.

- **Límite de QoS mín.**

  Representa la latencia de una carga de trabajo causada por la configuración de rendimiento mínimo de calidad de servicio (esperado) asignada a otras cargas de trabajo. Si el valor mínimo de calidad de servicio establecido en ciertas cargas de trabajo utiliza la mayoría del ancho de banda para garantizar el rendimiento prometido, se aceleran otras cargas de trabajo y muestran más latencia.

- **Interconexión en cluster**

  Representa los cables y los adaptadores que conectan de forma física los nodos en clúster. Si el componente de interconexión del clúster es objeto de disputa, significa que un tiempo de espera elevado para las solicitudes de I/o en la interconexión del clúster está afectando a la latencia de una o más cargas de trabajo.

- **Procesamiento de datos**

  Representa el componente de software del clúster involucrado en el procesamiento de I/o entre el clúster y el agregado de almacenamiento que contiene la carga de trabajo. Es posible que el nodo que gestiona el procesamiento de datos haya cambiado desde que se detectó el evento. Si el componente de procesamiento de datos es objeto de disputa, significa que un uso elevado en el nodo de procesamiento de datos está afectando a la latencia de una o más cargas de trabajo.

- **Activación de volumen**

  Representa el proceso que realiza un seguimiento del uso de todos los volúmenes activos. En entornos de gran tamaño en los que hay más de 1000 volúmenes activos, este proceso realiza un seguimiento de la cantidad de volúmenes críticos que se necesitan para acceder a los recursos a través del nodo al mismo tiempo. Cuando el número de volúmenes activos simultáneos supera el umbral máximo recomendado, algunos de los volúmenes no críticos experimentan latencia según se identifica aquí.

- **Recursos de MetroCluster**

  Representa los recursos de MetroCluster, incluidos NVRAM y los vínculos interswitch (ISL), que se usan para reflejar datos entre los clústeres de una configuración de MetroCluster. Si el componente MetroCluster es objeto de disputa, significa que el alto rendimiento de escritura de las cargas de trabajo del clúster local o un problema de estado del enlace afectan a la latencia de una o más cargas de trabajo del clúster local. Si el clúster no está en una configuración MetroCluster, este icono no se muestra.

- **Operaciones globales o agregados de SSD**

  Representa el agregado de almacenamiento en el que se ejecutan las cargas de trabajo. Si el componente de agregado es objeto de disputa, significa que un uso elevado en el agregado está afectando a la latencia de una o más cargas de trabajo. Un agregado está formado por todos los HDD, o una combinación de HDD y SSD (un agregado de Flash Pool), o una combinación de HDD y un nivel de cloud (un agregado de FabricPool). Un «agregado SD» está compuesto por todos los SSD (un agregado all-flash) o una combinación de SSD y un nivel de cloud (un agregado de FabricPool).

- **Latencia de cloud**

  Representa el componente de software del clúster involucrado en el procesamiento de I/o entre el clúster y el nivel de cloud en el que se almacenan los datos del usuario. Si el componente de latencia del cloud es objeto de disputa, significa que una gran cantidad de lecturas de volúmenes que están alojados en el nivel de cloud están afectando a la latencia de una o más cargas de trabajo.

- **SnapMirror** sincronizado

  Representa el componente de software del clúster involucrado en la replicación de datos de usuario del volumen primario al secundario en una relación de SnapMirror Synchronous. Si el componente Sync SnapMirror es objeto de disputa, significa que la actividad de las operaciones de SnapMirror Synchronous está afectando a la latencia de una o más cargas de trabajo.

Funciones de las cargas de trabajo involucradas en un evento de rendimiento

Unified Manager utiliza roles para identificar la implicación de una carga de trabajo en un evento de rendimiento. Entre las funciones se encuentran las víctimas, los agresores y los tiburones. Una carga de trabajo definida por el usuario puede ser víctima, abusador o tiburón al mismo tiempo.

Función	Descripción
Víctima	Carga de trabajo definida por el usuario cuyo rendimiento ha disminuido debido a otras cargas de trabajo, denominadas verdugos, que usan en exceso un componente de clúster. Solo las cargas de trabajo definidas por el usuario se identifican como víctimas. Unified Manager identifica las cargas de trabajo víctimas en función de su desviación en la latencia, donde la latencia real, durante un evento, ha aumentado considerablemente de su previsión de latencia (rango esperado).
Matón	Una carga de trabajo definida por el usuario o el sistema cuyo uso excesivo de un componente del clúster ha provocado la reducción del rendimiento de otras cargas de trabajo denominadas víctimas. Unified Manager identifica cargas de trabajo problemáticas según su desviación en el uso de un componente del clúster, donde el uso real, durante un evento, ha aumentado considerablemente desde su rango de uso esperado.
IBM	Carga de trabajo definida por el usuario con el mayor uso de un componente del clúster en comparación con todas las cargas de trabajo involucradas en un evento. Unified Manager identifica cargas de trabajo de tiburón en función de su uso de un componente de clúster durante un evento.

Las cargas de trabajo de un clúster pueden compartir muchos de los componentes del clúster, como agregados y la CPU, para el procesamiento de datos y red. Cuando una carga de trabajo, como un volumen, aumenta el uso de un componente del clúster hasta el punto de que el componente no puede satisfacer de forma eficiente las demandas de las cargas de trabajo, el componente es objeto de disputa. La carga de trabajo que utiliza en exceso un componente del clúster es un problema. Las otras cargas de trabajo que comparten esos componentes y cuyo rendimiento afecta al abusador son las víctimas. La actividad de las cargas de trabajo definidas por el sistema, como la deduplicación o las copias Snapshot, también puede convertirse en «bullying».

Cuando Unified Manager detecta un evento, identifica todas las cargas de trabajo y los componentes del clúster implicados, incluidas las cargas de trabajo abusivas que causaron el evento, el componente de clúster que está en disputa y las cargas de trabajo víctimas cuyo rendimiento ha disminuido debido al aumento de la actividad de las cargas de trabajo abusivas.

Si Unified Manager no puede identificar las cargas de trabajo abusivas, solo envía alertas sobre las cargas de trabajo víctimas y el componente de clúster correspondiente.

Unified Manager puede identificar cargas de trabajo que son víctimas de cargas de trabajo abusivas y, además, identificar cuándo esas mismas cargas de trabajo pasan a ser cargas de trabajo abusivas. Una carga de trabajo puede ser un problema para sí misma. Por ejemplo, una carga de trabajo de alto rendimiento que está acelerando por el límite de un grupo de políticas provoca que se aceleren todas las cargas de trabajo del grupo de políticas, incluso la propia. Una carga de trabajo que sea intimidatada o víctima en un evento de

rendimiento continuo puede cambiar su función o ya no ser participante en el evento.