Recogida de datos y supervisión del rendimiento de carga de trabajo

Active IQ Unified Manager 9.13

NetApp
December 18, 2023

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:

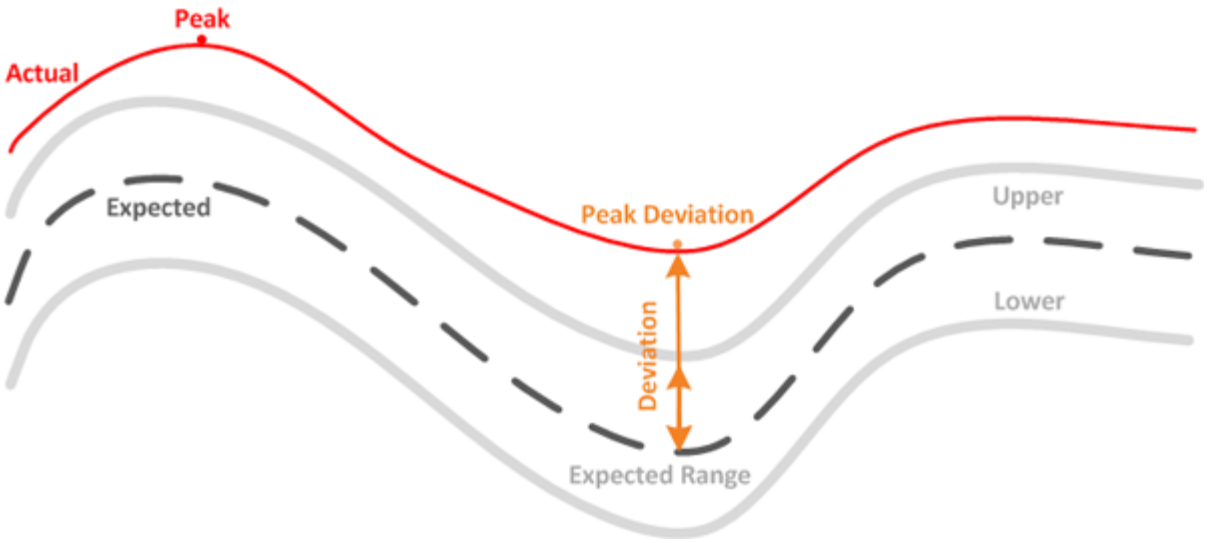
  - Eficiencia del almacenamiento, como la deduplicación
  - Estado del disco, que incluye la reconstrucción de RAID, el filtrado de discos, etc.
  - La replicación de datos, como las copias SnapMirror
  - Actividades de gestión
  - Estado del sistema de archivos, que incluye diversas actividades de WAFL
  - 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

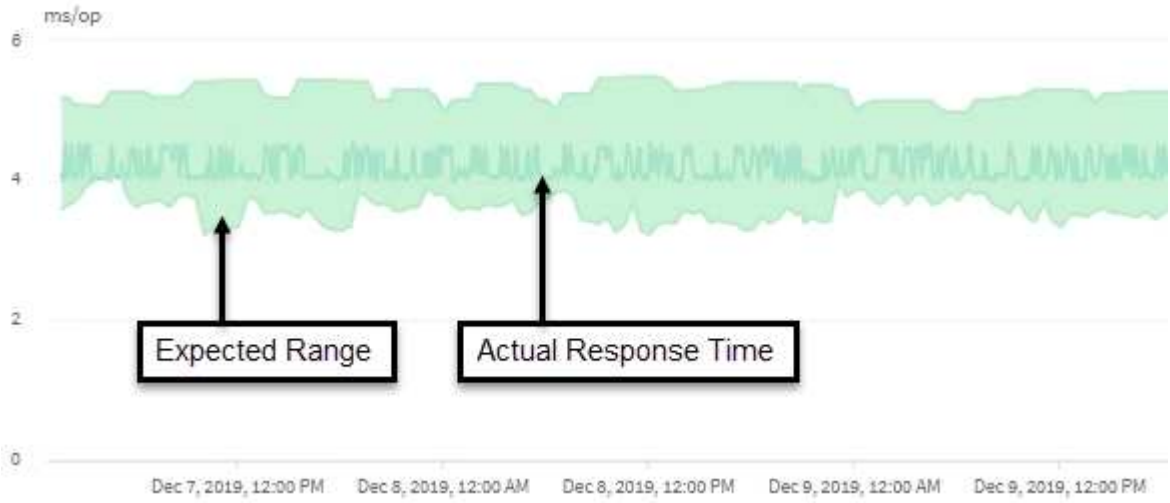
La latencia de las cargas de trabajo (tiempo de respuesta) es el tiempo que tarda un volumen en un clúster en responder a las solicitudes de I/o de las aplicaciones cliente. Unified Manager utiliza la latencia para detectar eventos de rendimiento y alertarle de ellos.

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.

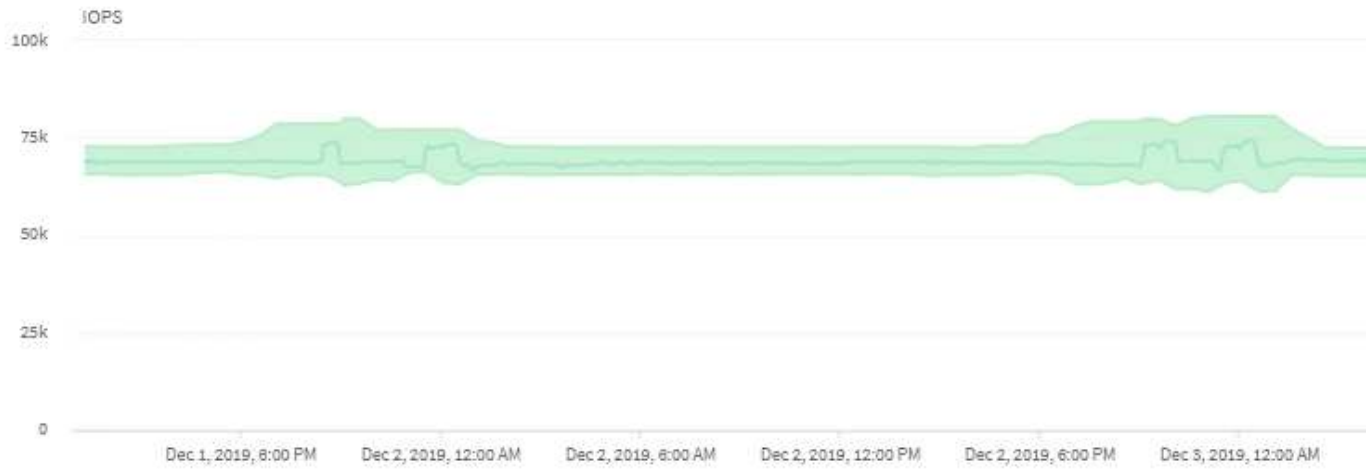
Puede haber vacíos en la línea azul si Unified Manager no pudo recopilar datos. Esto puede ocurrir porque el clúster o el volumen no se pudo acceder, Unified Manager se apagó durante esa hora o la recogida fue más prolongada que el período de recogida de 5 minutos.

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.

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.

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.

Puede ver el rendimiento de cualquiera de los clústeres de una configuración de MetroCluster analizando las

cargas de trabajo de los LUN y volúmenes correspondientes desde las siguientes pantallas. Puede filtrar los resultados por el clúster. Desde el panel de navegación de la izquierda:

- **Almacenamiento > clústeres > rendimiento: Vista todos los clústeres**. Consulte
- **Almacenamiento > volúmenes > rendimiento: Vista todos los volúmenes**.
- **Almacenamiento > LUN > rendimiento: Vista de todas las LUN**.
- **Análisis de cargas de trabajo > todas las cargas de trabajo**

Información relacionada

"Análisis y notificación de eventos de rendimiento"

"Análisis de eventos de rendimiento para una configuración de MetroCluster"

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

"Identificar las cargas de trabajo de las víctimas involucradas en un evento de rendimiento"

"Identificar las cargas de trabajo abusivas que están involucradas en un evento de rendimiento"

"Identificar las cargas de trabajo de tiburones en un evento de rendimiento"