Raccolta di dati e monitoraggio delle performance dei carichi di lavoro

Active IQ Unified Manager 9.7

NetApp
April 17, 2024

Sommario

Raccolta di dati e monitoraggio delle performance dei carichi di lavoro

Unified Manager raccoglie e analizza l'attività dei carichi di lavoro ogni 5 minuti per identificare gli eventi relativi alle performance e rileva le modifiche alla configurazione ogni 15 minuti. Conserva un massimo di 30 giorni di dati storici relativi alle performance e agli eventi di 5 minuti e utilizza questi dati per prevedere l'intervallo di latenza previsto per tutti i carichi di lavoro monitorati.

Unified Manager deve raccogliere un minimo di 3 giorni di attività del carico di lavoro prima che possa iniziare l'analisi e prima che la previsione di latenza per il tempo di risposta i/o possa essere visualizzata nella pagina workload Analysis e nella pagina Event Details. Durante la raccolta di questa attività, la previsione della latenza non visualizza tutte le modifiche che si verificano dall'attività del carico di lavoro. Dopo aver raccolto 3 giorni di attività, Unified Manager regola la previsione di latenza ogni 24 ore alle 12:00, per riflettere le modifiche dell'attività del carico di lavoro e stabilire una soglia di performance dinamica più precisa.

Durante i primi 4 giorni in cui Unified Manager sta monitorando un carico di lavoro, se sono trascorse più di 24 ore dall'ultima raccolta di dati, i grafici di latenza non visualizzano la previsione di latenza per quel carico di lavoro. Gli eventi rilevati prima dell'ultima raccolta sono ancora disponibili.

> L'ora legale (DST) modifica l'ora del sistema, che modifica la previsione di latenza delle statistiche delle performance per i carichi di lavoro monitorati. Unified Manager inizia immediatamente a correggere la previsione di latenza, che richiede circa 15 giorni per essere completata. Durante questo periodo di tempo è possibile continuare a utilizzare Unified Manager, ma poiché Unified Manager utilizza la previsione della latenza per rilevare eventi dinamici, alcuni eventi potrebbero non essere precisi. Gli eventi rilevati prima del cambiamento di orario non vengono influenzati.

Tipi di workload monitorati da Unified Manager

È possibile utilizzare Unified Manager per monitorare le performance di due tipi di carichi di lavoro: Definiti dall'utente e definiti dal sistema.

- ***workload definiti dall'utente***

  Il throughput di i/o dalle applicazioni al cluster. Si tratta di processi coinvolti nelle richieste di lettura e scrittura. Un volume, LUN, condivisione NFS, condivisione SMB/CIFS e un carico di lavoro è un carico di lavoro definito dall'utente.

  > Unified Manager monitora solo l'attività del carico di lavoro sul cluster. Non esegue il monitoraggio delle applicazioni, dei client o dei percorsi tra le applicazioni e il cluster.

  Se una o più delle seguenti affermazioni relative a un carico di lavoro sono vere, non possono essere monitorate da Unified Manager:

  - Si tratta di una copia di protezione dei dati (DP) in modalità di sola lettura. Tenere presente che i volumi DP vengono monitorati per il traffico generato dall'utente.
  - Si tratta di un volume infinito.
  - Si tratta di un clone dei dati offline.

    - Si tratta di un volume mirrorato in una configurazione MetroCluster.

- ***workload definiti dal sistema***

  I processi interni legati all'efficienza dello storage, alla replica dei dati e allo stato del sistema, tra cui:

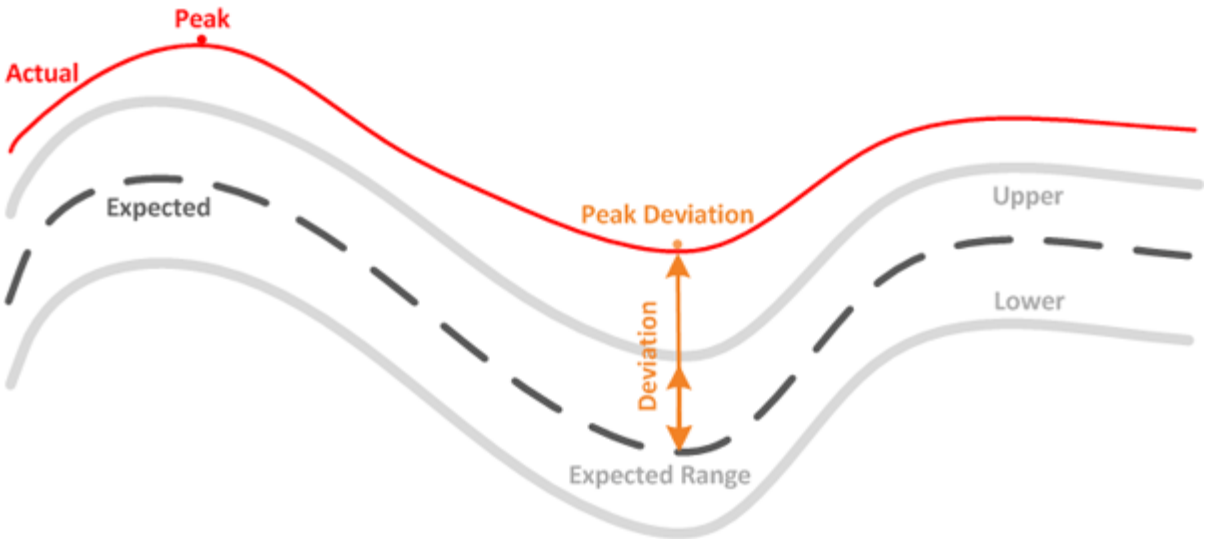
    - Efficienza dello storage, come la deduplica
    - Integrità del disco, che include la ricostruzione RAID, lo scrubbing del disco e così via
    - Replica dei dati, ad esempio le copie SnapMirror
    - Attività di gestione
    - Integrità del file system, che include varie attività di WAFL
    - Scanner del file system, come ad esempio la scansione WAFL
    - Offload delle copie, ad esempio operazioni di efficienza dello storage offload da host VMware
    - Stato del sistema, ad esempio spostamenti di volumi, compressione dei dati e così via
    - Volumi non monitorati

I dati sulle performance per i carichi di lavoro definiti dal sistema vengono visualizzati nella GUI solo quando il componente del cluster utilizzato da questi carichi di lavoro è in conflitto. Ad esempio, non è possibile cercare il nome di un carico di lavoro definito dal sistema per visualizzarne i dati sulle prestazioni nella GUI.

Valori di misurazione delle performance del carico di lavoro

Unified Manager misura le performance dei carichi di lavoro su un cluster in base a valori statistici storici e previsti, che formano la previsione di latenza dei valori per i carichi di lavoro. Confronta i valori statistici effettivi del carico di lavoro con la previsione di latenza per determinare quando le performance del carico di lavoro sono troppo alte o troppo basse. Un carico di lavoro che non funziona come previsto attiva un evento di performance dinamica per la notifica.

Nella seguente illustrazione, il valore effettivo, in rosso, rappresenta le statistiche effettive delle performance nel periodo di tempo. Il valore effettivo ha superato la soglia di performance, che è il limite superiore della previsione di latenza. Il picco è il valore effettivo più alto nell'intervallo di tempo. La deviazione misura la variazione tra i valori previsti (la previsione) e i valori effettivi, mentre la deviazione di picco indica la variazione maggiore tra i valori attesi e quelli effettivi.



La seguente tabella elenca i valori di misurazione delle performance del carico di lavoro.

Misurazione	Descrizione
Attività	La percentuale del limite di QoS utilizzato dai carichi di lavoro nel gruppo di criteri. Se Unified Manager rileva una modifica a un gruppo di criteri, ad esempio l'aggiunta o la rimozione di un volume o la modifica del limite di QoS, i valori effettivi e previsti potrebbero superare il 100% del limite impostato. Se un valore supera il 100% del limite impostato, viene visualizzato come >100%. Se un valore è inferiore all'1% del limite impostato, viene visualizzato come <1%.
Effettivo	Il valore misurato delle performance in un momento specifico per un determinato carico di lavoro.
Deviazione	Il cambiamento tra i valori previsti e quelli effettivi. Si tratta del rapporto tra il valore effettivo meno il valore previsto e il valore superiore dell'intervallo previsto meno il valore previsto. Un valore di deviazione negativo indica che le performance del carico di lavoro sono inferiori al previsto, mentre un valore di deviazione positivo indica che le performance del carico di lavoro sono superiori al previsto.

Misurazione	Descrizione
Previsto	I valori previsti si basano sull'analisi dei dati storici delle performance per un determinato carico di lavoro. Unified Manager analizza questi valori statistici per determinare l'intervallo previsto (previsione di latenza) dei valori.
Previsione di latenza (intervallo previsto)	La previsione di latenza è una previsione dei valori di performance superiori e inferiori previsti in un momento specifico. Per la latenza del carico di lavoro, i valori superiori costituiscono la soglia di performance. Quando il valore effettivo supera la soglia di performance, Unified Manager attiva un evento di performance dinamico.
Picco	Il valore massimo misurato in un periodo di tempo.
Deviazione di picco	Il valore di deviazione massimo misurato in un periodo di tempo.
Profondità della coda	Il numero di richieste i/o in sospeso che sono in attesa sul componente di interconnessione.
Utilizzo	Per l'elaborazione di rete, l'elaborazione dei dati e i componenti aggregati, la percentuale di tempo occupato per completare le operazioni dei carichi di lavoro in un determinato periodo di tempo. Ad esempio, la percentuale di tempo in cui i componenti di elaborazione dati o di elaborazione di rete elaborano una richiesta di i/o o un aggregato deve soddisfare una richiesta di lettura o scrittura.
Throughput in scrittura	La quantità di throughput in scrittura, espressa in megabyte al secondo (MB/s), dai carichi di lavoro su un cluster locale al cluster partner in una configurazione MetroCluster.

Qual è la gamma di performance prevista

La previsione di latenza è una previsione dei valori di performance superiori e inferiori previsti in un momento specifico. Per la latenza del carico di lavoro, i valori superiori costituiscono la soglia di performance. Quando il valore effettivo supera la soglia di performance, Unified Manager attiva un evento di performance dinamico.

Ad esempio, durante le normali ore di lavoro tra le 9:00 e alle 17:00, la maggior parte dei dipendenti potrebbe controllare la posta elettronica tra le 9:00 e alle 10:30 L'aumento della domanda sui server di posta elettronica comporta un aumento dell'attività dei carichi di lavoro sullo storage back-end durante questo periodo. I dipendenti potrebbero notare tempi di risposta lenti dai propri client di posta elettronica.

Durante l'ora di pranzo tra le 12:00 e alle 13:00 e alla fine della giornata lavorativa dopo le 17:00, la maggior parte dei dipendenti è probabilmente lontana dai computer. La domanda sui server di posta elettronica in genere diminuisce, diminuendo anche la domanda sullo storage back-end. In alternativa, potrebbero essere pianificate operazioni di carico di lavoro, come backup dello storage o scansione virus, che iniziano dopo le 17:00 e aumentare l'attività sullo storage back-end.

Nel corso di diversi giorni, l'aumento e la diminuzione dell'attività del carico di lavoro determina l'intervallo previsto (previsione di latenza) dell'attività, con limiti superiori e inferiori per un carico di lavoro. Quando l'attività effettiva del carico di lavoro di un oggetto si trova al di fuori dei limiti superiori o inferiori e rimane al di fuori dei limiti per un certo periodo di tempo, ciò potrebbe indicare che l'oggetto è stato utilizzato in eccesso o sottoutilizzato.

Come si forma la previsione di latenza

Unified Manager deve raccogliere un minimo di 3 giorni di attività del carico di lavoro prima che possa iniziare l'analisi e prima che la previsione di latenza per il tempo di risposta i/o possa essere visualizzata nella GUI. La raccolta dati minima richiesta non tiene conto di tutte le modifiche che si verificano dall'attività del carico di lavoro. Dopo aver raccolto i primi 3 giorni di attività, Unified Manager regola la previsione di latenza ogni 24 ore alle 12:00 riflettere le modifiche dell'attività del carico di lavoro e stabilire una soglia di performance dinamica più precisa.

L'ora legale (DST) modifica l'ora del sistema, che modifica la previsione di latenza delle statistiche delle performance per i carichi di lavoro monitorati. Unified Manager inizia immediatamente a correggere la previsione di latenza, che richiede circa 15 giorni per essere completata. Durante questo periodo di tempo è possibile continuare a utilizzare Unified Manager, ma poiché Unified Manager utilizza la previsione della latenza per rilevare eventi dinamici, alcuni eventi potrebbero non essere precisi. Gli eventi rilevati prima del cambiamento di orario non vengono influenzati.

Come viene utilizzata la previsione di latenza nell'analisi delle performance

Unified Manager utilizza la previsione della latenza per rappresentare la tipica attività di latenza i/o (tempo di risposta) per i carichi di lavoro monitorati. Ti avvisa quando la latenza effettiva per un carico di lavoro supera i limiti superiori della previsione di latenza, che attiva un evento di performance dinamica, in modo da poter analizzare il problema delle performance e intraprendere azioni correttive per risolverlo.

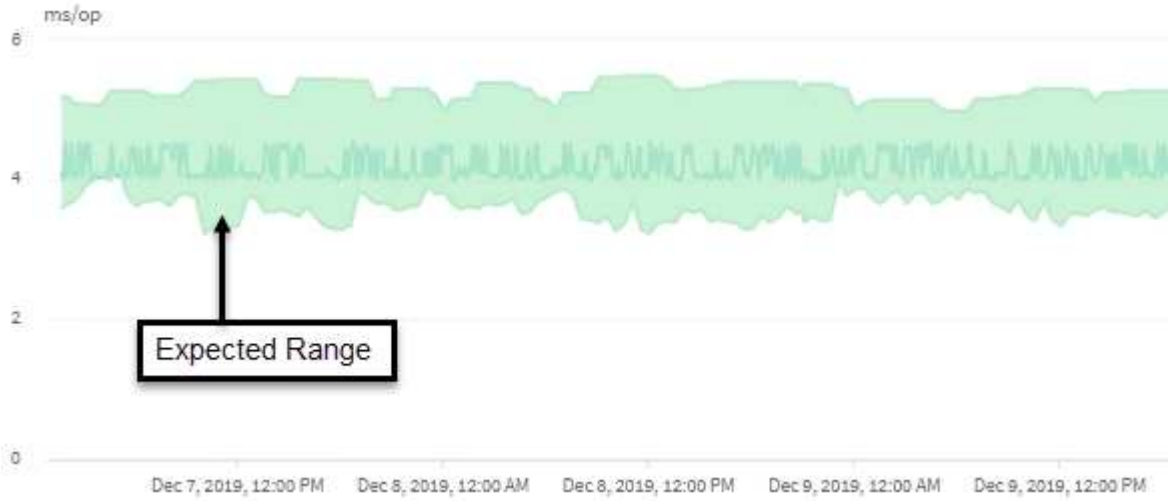
La previsione della latenza definisce la linea di base delle performance per il carico di lavoro. Nel corso del tempo, Unified Manager apprende dalle precedenti misurazioni delle performance per prevedere i livelli di performance e attività previsti per il carico di lavoro. Il limite superiore dell'intervallo previsto stabilisce la soglia di performance dinamica. Unified Manager utilizza la linea di base per determinare quando la latenza effettiva è superiore o inferiore a una soglia o al di fuori dei limiti previsti. Il confronto tra i valori effettivi e quelli previsti crea un profilo di performance per il carico di lavoro.

Quando la latenza effettiva per un carico di lavoro supera la soglia di performance dinamica, a causa di un conflitto su un componente del cluster, la latenza è elevata e il carico di lavoro funziona più lentamente del previsto. Anche le performance di altri carichi di lavoro che condividono gli stessi componenti del cluster potrebbero essere più lente del previsto.

Unified Manager analizza l'evento di superamento della soglia e determina se l'attività è un evento di performance. Se l'elevata attività del carico di lavoro rimane costante per un lungo periodo di tempo, ad

esempio diverse ore, Unified Manager considera l'attività normale e regola dinamicamente la previsione di latenza per formare la nuova soglia di performance dinamica.

Alcuni carichi di lavoro potrebbero avere un'attività costantemente bassa, dove la previsione di latenza per la latenza non ha un elevato tasso di cambiamento nel tempo. Per ridurre al minimo il numero di eventi durante l'analisi degli eventi delle performance, Unified Manager attiva un evento solo per volumi a bassa attività le cui operazioni e latenze sono molto più elevate del previsto.



In questo esempio, la latenza per un volume ha una previsione di latenza, in grigio, di 3.5 millisecondi per operazione (ms/op) al minimo e di 5.5 ms/op al massimo. Se la latenza effettiva, in blu, aumenta improvvisamente a 10 ms/op, a causa di un picco intermittente nel traffico di rete o di un conflitto su un componente del cluster, supera la previsione di latenza e supera la soglia di performance dinamica.

Quando il traffico di rete è diminuito o il componente del cluster non è più in conflitto, la latenza ritorna entro la previsione di latenza. Se la latenza rimane pari o superiore a 10 ms/op per un lungo periodo di tempo, potrebbe essere necessario intraprendere un'azione correttiva per risolvere l'evento.

Come Unified Manager utilizza la latenza dei workload per identificare i problemi di performance

La latenza del carico di lavoro (tempo di risposta) è il tempo necessario a un volume di un cluster per rispondere alle richieste di i/o provenienti dalle applicazioni client. Unified Manager utilizza la latenza per rilevare e avvisare gli utenti in caso di eventi relativi alle performance.

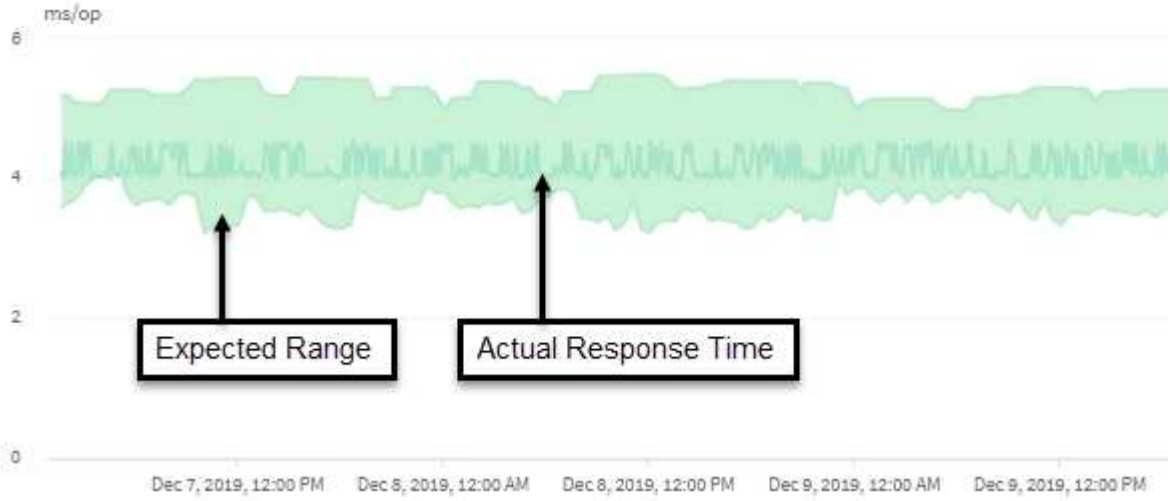
Un'elevata latenza significa che le richieste dalle applicazioni a un volume su un cluster richiedono più tempo del solito. La causa dell'elevata latenza potrebbe essere il cluster stesso, a causa di conflitti su uno o più componenti del cluster. L'elevata latenza potrebbe essere causata anche da problemi esterni al cluster, come colli di bottiglia della rete, problemi con il client che ospita le applicazioni o problemi con le applicazioni stesse.

Unified Manager monitora solo l'attività del carico di lavoro sul cluster. Non esegue il monitoraggio delle applicazioni, dei client o dei percorsi tra le applicazioni e il cluster.

Anche le operazioni sul cluster, come la creazione di backup o l'esecuzione della deduplica, che aumentano la domanda di componenti del cluster condivisi da altri carichi di lavoro possono contribuire a un'elevata latenza. Se la latenza effettiva supera la soglia di performance dinamica dell'intervallo previsto (previsione di latenza), Unified Manager analizza l'evento per determinare se si tratta di un evento di performance che potrebbe

essere necessario risolvere. La latenza viene misurata in millisecondi per operazione (ms/op).

Nel grafico latenza totale della pagina analisi del carico di lavoro, è possibile visualizzare un'analisi delle statistiche di latenza per vedere come l'attività dei singoli processi, come le richieste di lettura e scrittura, si confronta con le statistiche di latenza complessive. Il confronto consente di determinare quali operazioni hanno l'attività più elevata o se operazioni specifiche hanno attività anomale che influiscono sulla latenza di un volume. Quando si analizzano gli eventi delle performance, è possibile utilizzare le statistiche di latenza per determinare se un evento è stato causato da un problema nel cluster. È inoltre possibile identificare le attività specifiche del carico di lavoro o i componenti del cluster coinvolti nell'evento.



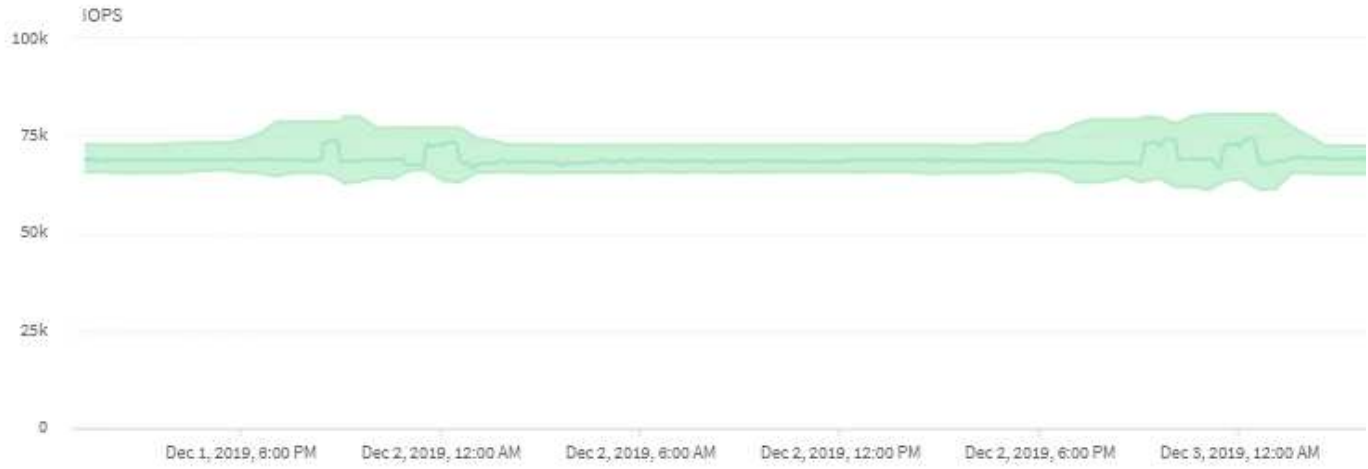
Questo esempio mostra il grafico della latenza . L'attività del tempo di risposta effettivo (latenza) è una linea blu e la previsione di latenza (intervallo previsto) è verde.

Se Unified Manager non è in grado di raccogliere i dati, la linea blu può presentare delle lacune. Ciò può verificarsi perché il cluster o il volume non era raggiungibile, Unified Manager è stato disattivato durante tale periodo o la raccolta richiede più tempo del periodo di raccolta di 5 minuti.

In che modo le operazioni del cluster possono influire sulla latenza del carico di lavoro

Le operazioni (IOPS) rappresentano l'attività di tutti i carichi di lavoro definiti dall'utente e dal sistema su un cluster. Le statistiche IOPS consentono di determinare se i processi del cluster, come l'esecuzione di backup o la deduplica, influiscono sulla latenza del carico di lavoro (tempo di risposta) o potrebbero aver causato o contribuito a un evento di performance.

Quando si analizzano gli eventi relativi alle performance, è possibile utilizzare le statistiche IOPS per determinare se un evento relativo alle performance è stato causato da un problema nel cluster. È possibile identificare le attività specifiche dei carichi di lavoro che potrebbero aver contribuito in maniera determinante all'evento delle performance. Gli IOPS vengono misurati in operazioni al secondo (Ops/sec).



Questo esempio mostra il grafico IOPS. Le statistiche effettive delle operazioni sono una linea blu e le previsioni IOPS delle statistiche delle operazioni sono verdi.

In alcuni casi in cui un cluster è sovraccarico, Unified Manager potrebbe visualizzare il messaggio `Data collection is taking too long on Cluster cluster_name`. Ciò significa che non sono state raccolte statistiche sufficienti per l'analisi di Unified Manager. È necessario ridurre le risorse utilizzate dal cluster in modo da poter raccogliere le statistiche.

Monitoraggio delle performance delle configurazioni MetroCluster

Unified Manager consente di monitorare il throughput di scrittura tra i cluster in una configurazione MetroCluster per identificare i carichi di lavoro con un'elevata quantità di throughput in scrittura. Se questi carichi di lavoro dalle performance elevate causano elevati tempi di risposta i/o per altri volumi nel cluster locale, Unified Manager attiva gli eventi relativi alle performance per ricevere una notifica.

Quando un cluster locale in una configurazione MetroCluster esegue il mirroring dei dati nel cluster partner, i dati vengono scritti nella NVRAM e quindi trasferiti attraverso i collegamenti interswitch (ISL) agli aggregati remoti. Unified Manager analizza la NVRAM per identificare i carichi di lavoro il cui throughput di scrittura elevato sta utilizzando la NVRAM in eccesso, mettendo la NVRAM in conflitto.

I carichi di lavoro la cui deviazione nel tempo di risposta ha superato la soglia di performance sono denominati *vittime* e i carichi di lavoro la cui deviazione nel throughput di scrittura nella NVRAM è superiore al solito, causando il conflitto, sono denominati *bullies*. Poiché solo le richieste di scrittura vengono mirrorate al cluster partner, Unified Manager non analizza il throughput in lettura.

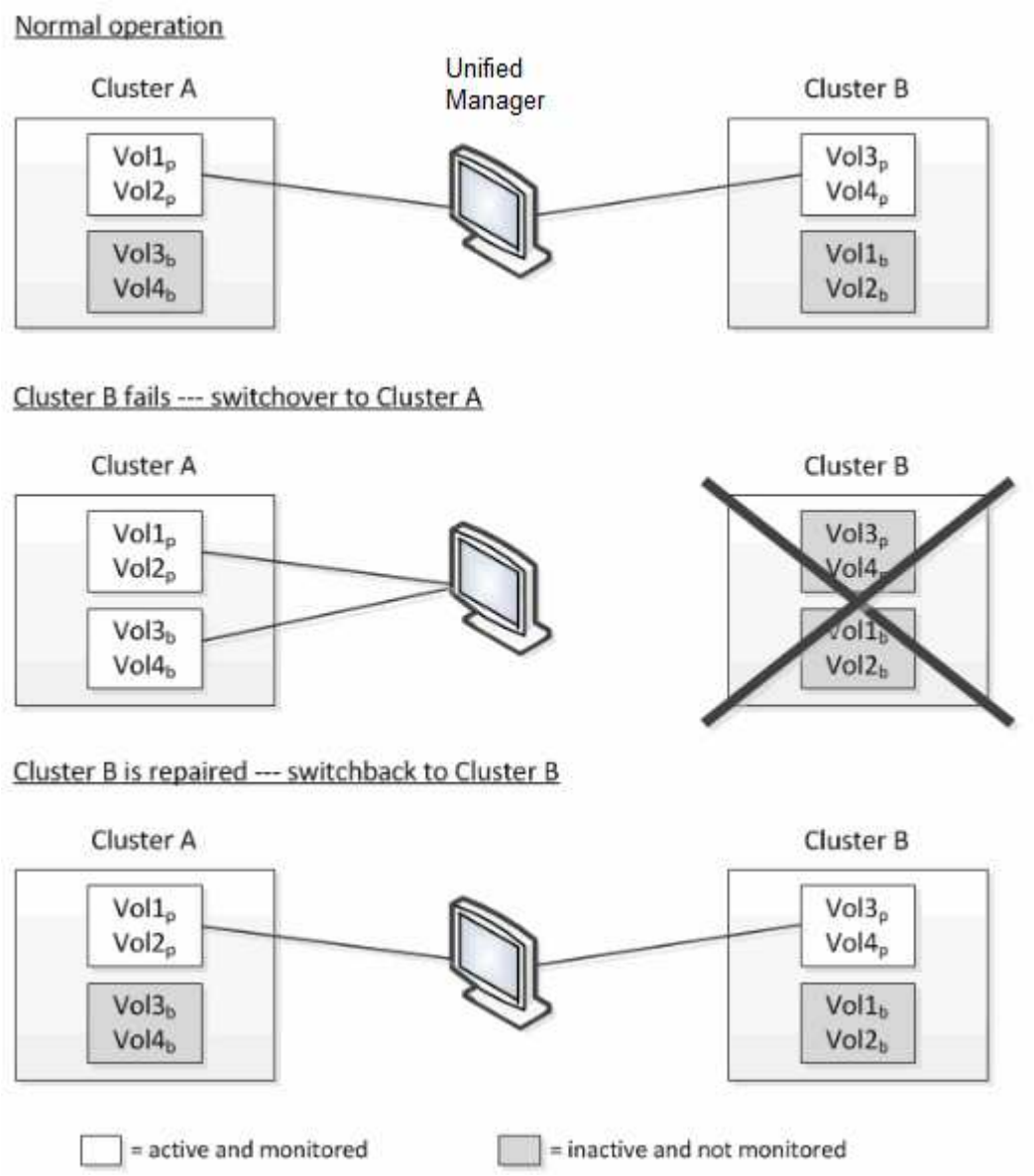
Unified Manager tratta i cluster in una configurazione MetroCluster come singoli cluster. Non distingue i cluster che sono partner o correlano il throughput di scrittura da ciascun cluster.

Comportamento del volume durante lo switchover e lo switchback

Gli eventi che attivano uno switchover o uno switchback causano lo spostamento dei volumi attivi da un cluster all'altro nel gruppo di disaster recovery. I volumi sul cluster attivi e che forniscono dati ai client vengono arrestati e i volumi sull'altro cluster vengono

attivati e iniziano a servire i dati. Unified Manager monitora solo i volumi attivi e in esecuzione.

Poiché i volumi vengono spostati da un cluster all'altro, si consiglia di monitorare entrambi i cluster. Una singola istanza di Unified Manager può monitorare entrambi i cluster in una configurazione MetroCluster, ma a volte la distanza tra le due posizioni richiede l'utilizzo di due istanze di Unified Manager per monitorare entrambi i cluster. La figura seguente mostra una singola istanza di Unified Manager:



I volumi con p nei loro nomi indicano i volumi primari e i volumi con b nei loro nomi sono volumi di backup mirrorati creati da SnapMirror.

Durante il normale funzionamento:

- Il cluster A ha due volumi attivi: Vol1p e Vol2p.
- Il cluster B ha due volumi attivi: Vol3p e Vol4p.
- Il cluster A ha due volumi inattivi: Vol3b e Vol4b.
- Il cluster B ha due volumi inattivi: Vol1b e Vol2b.

Unified Manager raccoglie le informazioni relative a ciascuno dei volumi attivi (statistiche, eventi e così via). Le statistiche Vol1p e Vol2p vengono raccolte dal cluster A e le statistiche Vol3p e Vol4p vengono raccolte dal cluster B.

Dopo un guasto catastrofico che causa lo switchover dei volumi attivi dal cluster B al cluster A:

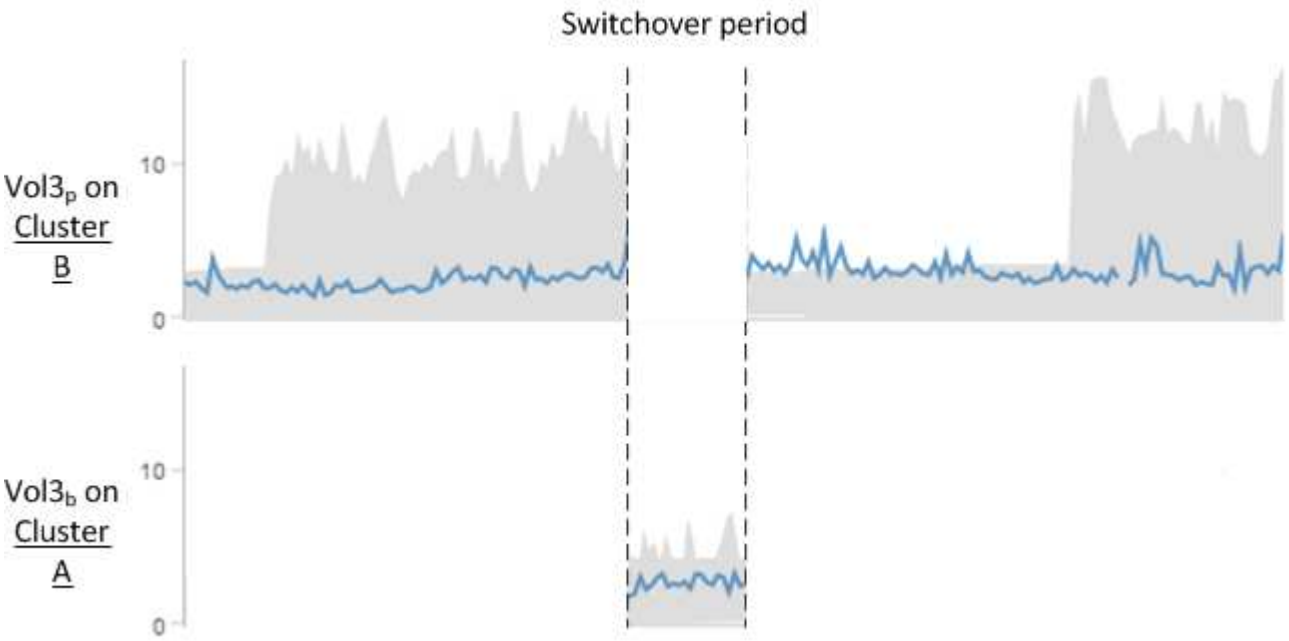
- Il cluster A ha quattro volumi attivi: Vol1p, Vol2p, Vol3b e Vol4b.
- Il cluster B ha quattro volumi inattivi: Vol3p, Vol4p, Vol1b e Vol2b.

Come durante il normale funzionamento, Unified Manager raccoglie le informazioni relative a ciascuno dei volumi attivi. Tuttavia, in questo caso, le statistiche Vol1p e Vol2p vengono raccolte dal cluster A, mentre le statistiche Vol3b e Vol4b vengono raccolte anche dal cluster A.

Si noti che Vol3p e Vol3b non sono gli stessi volumi, perché si trovano su cluster diversi. Le informazioni di Unified Manager per Vol3p non sono le stesse di Vol3b:

- Durante il passaggio al cluster A, le statistiche e gli eventi di Vol3p non sono visibili.
- Al primo passaggio, Vol3b sembra un nuovo volume senza informazioni storiche.

Quando il cluster B viene riparato e viene eseguito uno switchback, il Vol3p viene nuovamente attivato sul cluster B, con le statistiche storiche e un intervallo di statistiche per il periodo durante lo switchover. Vol3b non è visualizzabile dal cluster A fino a quando non si verifica un altro switchover:



- I volumi MetroCluster inattivi, ad esempio Vol3b sul cluster A dopo lo switchback, vengono identificati con il messaggio “questo volume è stato cancellato”. Il volume non viene effettivamente eliminato, ma non viene attualmente monitorato da Unified Manager perché non è il volume attivo.
- Se un singolo Unified Manager sta monitorando entrambi i cluster in una configurazione MetroCluster, la ricerca del volume restituisce informazioni per il volume attivo in quel momento. Ad esempio, una ricerca di “Vol3” restituisce statistiche ed eventi per Vol3b sul cluster A se si è verificato uno switchover e Vol3 è diventato attivo sul cluster A.

Analisi e notifica degli eventi relativi alle performance

Gli eventi relativi alle performance avvisano l'utente in merito a problemi di performance i/o su un carico di lavoro causati da conflitti su un componente del cluster. Unified Manager analizza l'evento per identificare tutti i carichi di lavoro coinvolti, il componente in conflitto e se l'evento è ancora un problema che potrebbe essere necessario risolvere.

Unified Manager monitora la latenza di i/o (tempo di risposta) e gli IOPS (operazioni) per i volumi su un cluster. Quando altri carichi di lavoro utilizzano in eccesso un componente del cluster, ad esempio, il componente è in conflitto e non può funzionare a un livello ottimale per soddisfare le esigenze dei carichi di lavoro. Le performance di altri carichi di lavoro che utilizzano lo stesso componente potrebbero risentirne, causando un aumento delle latenze. Se la latenza supera la soglia dinamica delle performance, Unified Manager attiva un evento di performance per avvisare l'utente.

Analisi degli eventi

Unified Manager esegue le seguenti analisi, utilizzando i 15 giorni precedenti di statistiche sulle performance, per identificare i carichi di lavoro delle vittime, i carichi di lavoro ingombranti e il componente del cluster coinvolto in un evento:

- Identifica i carichi di lavoro delle vittime la cui latenza ha superato la soglia di performance dinamica, che è il limite superiore della previsione di latenza:
  - Per i volumi su aggregati HDD o Flash Pool (ibridi) (Tier locale), gli eventi vengono attivati solo quando la latenza è superiore a 5 millisecondi (ms) e gli IOPS sono superiori a 10 operazioni al secondo (Ops/sec).
  - Per i volumi su aggregati all-SSD o aggregati FabricPool (cloud Tier), gli eventi vengono attivati solo quando la latenza è superiore a 1 ms e gli IOPS sono superiori a 100 Ops/sec.
- Identifica il componente del cluster in conflitto.

  Se la latenza dei carichi di lavoro delle vittime nell'interconnessione del cluster è superiore a 1 ms, Unified Manager considera questa condizione come significativa e attiva un evento per l'interconnessione del cluster.

- Identifica i carichi di lavoro ingombranti che stanno utilizzando in eccesso il componente del cluster e che lo causano in conflitto.
- Classifica i carichi di lavoro coinvolti, in base alla loro deviazione nell'utilizzo o nell'attività di un componente del cluster, per determinare quali bulli hanno il cambiamento più elevato nell'utilizzo del componente del cluster e quali sono le vittime più interessate.

Un evento potrebbe verificarsi solo per un breve momento e poi correggersi una volta che il componente che sta utilizzando non è più in conflitto. Un evento continuo si verifica nuovamente per lo stesso componente del cluster entro un intervallo di cinque minuti e rimane nello stato attivo. Per gli eventi continui, Unified Manager attiva un avviso dopo aver rilevato lo stesso evento durante due intervalli di analisi consecutivi.

Quando un evento viene risolto, rimane disponibile in Unified Manager come parte della registrazione dei problemi di performance passati per un volume. Ogni evento ha un ID univoco che identifica il tipo di evento e i volumi, il cluster e i componenti del cluster coinvolti.

Un singolo volume può essere coinvolto in più eventi contemporaneamente.

Stato dell'evento

Gli eventi possono trovarsi in uno dei seguenti stati:

- **Attivo**

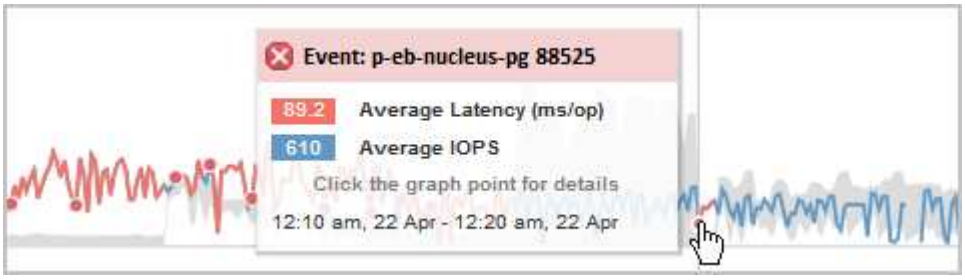
  Indica che l'evento di performance è attualmente attivo (nuovo o confermato). Il problema che causa l'evento non è stato risolto o non è stato risolto. Il contatore delle performance per l'oggetto storage rimane al di sopra della soglia di performance.

- **Obsoleto**

  Indica che l'evento non è più attivo. Il problema che ha causato l'evento è stato risolto o risolto. Il contatore delle performance per l'oggetto storage non è più al di sopra della soglia di performance.

Notifica degli eventi

Gli eventi vengono visualizzati nella pagina Dashboard e in molte altre pagine dell'interfaccia utente e gli avvisi relativi a tali eventi vengono inviati a indirizzi e-mail specifici. È possibile visualizzare informazioni di analisi dettagliate su un evento e ottenere suggerimenti per risolverlo nella pagina Dettagli evento e nella pagina analisi del carico di lavoro.



In questo esempio, un evento è indicato da un punto rosso (🔴) Sul grafico della latenza. Passando il cursore del mouse sul punto rosso viene visualizzata una finestra a comparsa con ulteriori dettagli sull'evento e opzioni per analizzarlo.

Interazione con gli eventi

Nella pagina Dettagli evento e nella pagina analisi del carico di lavoro, è possibile interagire con gli eventi nei seguenti modi:

- Spostando il mouse su un evento, viene visualizzato un messaggio che mostra l'ID dell'evento e la data e l'ora in cui è stato rilevato l'evento.

  Se sono presenti più eventi per lo stesso periodo di tempo, il messaggio mostra il numero di eventi.

- Facendo clic su un singolo evento viene visualizzata una finestra di dialogo che mostra informazioni più dettagliate sull'evento, inclusi i componenti del cluster coinvolti.

  Il componente in conflitto viene cerchiato ed evidenziato in rosso. È possibile fare clic sull'ID evento o su **Visualizza analisi completa** per visualizzare l'analisi completa nella pagina Dettagli evento. Se sono presenti più eventi per lo stesso periodo di tempo, la finestra di dialogo mostra i dettagli relativi ai tre eventi più recenti. È possibile fare clic su un ID evento per visualizzare l'analisi degli eventi nella pagina Dettagli evento.

In che modo Unified Manager determina l'impatto delle performance di un evento

Unified Manager utilizza la deviazione nell'attività, nell'utilizzo, nel throughput di scrittura, nell'utilizzo dei componenti del cluster o nella latenza di i/o (tempo di risposta) per un carico di lavoro per determinare il livello di impatto sulle performance del carico di lavoro. Queste informazioni determinano il ruolo di ciascun carico di lavoro nell'evento e il modo in cui sono classificati nella pagina Dettagli evento.

Unified Manager confronta gli ultimi valori analizzati per un carico di lavoro con l'intervallo previsto (previsione di latenza) dei valori. La differenza tra gli ultimi valori analizzati e l'intervallo di valori previsto identifica i carichi di lavoro le cui performance sono state maggiormente influenzate dall'evento.

Ad esempio, supponiamo che un cluster contenga due carichi di lavoro: Workload A e workload B. La previsione di latenza per il carico di lavoro A è di 5-10 millisecondi per operazione (ms/op) e la latenza effettiva è di solito di circa 7 ms/op. La latenza prevista per il carico di lavoro B è di 10-20 ms/op e la latenza effettiva è di solito di circa 15 ms/op. Entrambi i carichi di lavoro rientrano nella loro previsione di latenza. A causa del conflitto sul cluster, la latenza di entrambi i carichi di lavoro aumenta fino a 40 ms/op, superando la soglia di performance dinamica, che è il limite superiore della previsione di latenza, e attivando gli eventi. La deviazione nella latenza, dai valori previsti ai valori superiori alla soglia di performance, per il carico di lavoro A è di circa 33 ms/op e la deviazione per il carico di lavoro B è di circa 25 ms/op. La latenza di entrambi i carichi di lavoro è aumentata fino a 40 ms/op, ma il carico di lavoro A ha avuto un impatto maggiore sulle performance perché aveva una maggiore deviazione della latenza a 33 ms/op.

Nella pagina Dettagli evento, nella sezione Diagnosi del sistema, è possibile ordinare i carichi di lavoro in base alla loro deviazione nell'attività, nell'utilizzo o nel throughput per un componente del cluster. Puoi anche ordinare i workload in base alla latenza. Quando si seleziona un'opzione di ordinamento, Unified Manager analizza la deviazione nell'attività, nell'utilizzo, nel throughput o nella latenza dal momento in cui l'evento è stato rilevato dai valori previsti per determinare l'ordinamento dei carichi di lavoro. Per la latenza, i punti rossi (🔴) indicano un superamento della soglia di performance da parte di un carico di lavoro della vittima e il conseguente impatto sulla latenza. Ogni punto rosso indica un livello più elevato di deviazione nella latenza, che consente di identificare i carichi di lavoro delle vittime la cui latenza è stata maggiormente influenzata da un evento.

Componenti del cluster e perché possono essere in conflitto

È possibile identificare i problemi di performance del cluster quando un componente del cluster entra in conflitto. Le performance dei carichi di lavoro che utilizzano il componente rallentano e il loro tempo di risposta (latenza) per le richieste dei client aumenta, il che attiva un evento in Unified Manager.

Un componente in conflitto non può funzionare a un livello ottimale. Le sue performance sono diminuite e le performance di altri componenti e carichi di lavoro del cluster, denominati *vittime*, potrebbero avere una maggiore latenza. Per eliminare un componente dai conflitti, è necessario ridurre il carico di lavoro o aumentare la capacità di gestire più lavoro, in modo che le performance possano tornare ai livelli normali. Poiché Unified Manager raccoglie e analizza le performance dei carichi di lavoro in intervalli di cinque minuti, rileva solo quando un componente del cluster viene costantemente utilizzato in eccesso. I picchi transitori di utilizzo eccessivo che durano solo per una breve durata nell'intervallo di cinque minuti non vengono rilevati.

Ad esempio, un aggregato di storage potrebbe essere in conflitto perché uno o più carichi di lavoro su di esso sono in competizione per soddisfare le richieste di i/O. Altri carichi di lavoro sull'aggregato possono risentirne, causando una diminuzione delle performance. Per ridurre la quantità di attività sull'aggregato, è possibile eseguire diverse operazioni, ad esempio lo spostamento di uno o più carichi di lavoro in un aggregato o nodo

meno occupato, per ridurre la domanda complessiva del carico di lavoro sull'aggregato corrente. Per un gruppo di policy QoS, è possibile regolare il limite di throughput o spostare i carichi di lavoro in un gruppo di policy diverso, in modo che i carichi di lavoro non vengano più rallentati.

Unified Manager monitora i seguenti componenti del cluster per avvisare l'utente quando si trovano in conflitto:

- **Rete**

  Rappresenta il tempo di attesa delle richieste di i/o da parte dei protocolli di rete esterni sul cluster. Il tempo di attesa è il tempo impiegato in attesa del completamento delle transazioni "transfer ready" prima che il cluster possa rispondere a una richiesta di i/O. Se il componente di rete è in conflitto, significa che il tempo di attesa elevato a livello di protocollo influisce sulla latenza di uno o più carichi di lavoro.

- **Elaborazione di rete**

  Rappresenta il componente software del cluster coinvolto nell'elaborazione i/o tra il livello di protocollo e il cluster. Il nodo che gestisce l'elaborazione di rete potrebbe essere cambiato da quando è stato rilevato l'evento. Se il componente di elaborazione di rete è in conflitto, significa che un utilizzo elevato nel nodo di elaborazione di rete influisce sulla latenza di uno o più carichi di lavoro.

  Quando si utilizza un cluster All SAN Array in una configurazione Active-Active, il valore di latenza di elaborazione della rete viene visualizzato per entrambi i nodi, in modo da poter verificare che i nodi condividano il carico in maniera uguale.

- **QoS Limit Max**

  Rappresenta l'impostazione di throughput massimo (picco) del gruppo di criteri QoS (Quality of Service) dello storage assegnato al carico di lavoro. Se il componente del gruppo di policy è in conflitto, significa che tutti i carichi di lavoro nel gruppo di policy vengono rallentati dal limite di throughput impostato, il che influisce sulla latenza di uno o più di tali carichi di lavoro.

- **Limite QoS min**

  Rappresenta la latenza per un carico di lavoro causata dall'impostazione QoS throughput Minimum (previsto) assegnata ad altri carichi di lavoro. Se il valore minimo di QoS impostato su alcuni carichi di lavoro utilizza la maggior parte della larghezza di banda per garantire il throughput promesso, altri carichi di lavoro verranno rallentati e otterranno una maggiore latenza.

- **Interconnessione cluster**

  Rappresenta i cavi e gli adattatori con cui i nodi in cluster sono fisicamente connessi. Se il componente di interconnessione del cluster è in conflitto, significa che l'elevato tempo di attesa per le richieste di i/o dell'interconnessione del cluster influisce sulla latenza di uno o più carichi di lavoro.

- **Elaborazione dei dati**

  Rappresenta il componente software del cluster coinvolto nell'elaborazione i/o tra il cluster e l'aggregato di storage che contiene il carico di lavoro. Il nodo che gestisce l'elaborazione dei dati potrebbe essere cambiato da quando è stato rilevato l'evento. Se il componente di elaborazione dei dati è in conflitto, significa che un utilizzo elevato nel nodo di elaborazione dei dati influisce sulla latenza di uno o più carichi di lavoro.

- **Attivazione del volume**

  Rappresenta il processo che tiene traccia dell'utilizzo di tutti i volumi attivi. In ambienti di grandi dimensioni in cui sono attivi più di 1000 volumi, questo processo tiene traccia del numero di volumi critici necessari per

accedere alle risorse attraverso il nodo allo stesso tempo. Quando il numero di volumi attivi simultanei supera la soglia massima consigliata, alcuni volumi non critici sperimenteranno la latenza come indicato qui.

- **Risorse MetroCluster**

  Rappresenta le risorse MetroCluster, tra cui NVRAM e ISL (Interswitch link), utilizzate per eseguire il mirroring dei dati tra cluster in una configurazione MetroCluster. Se il componente MetroCluster è in conflitto, significa che un elevato throughput di scrittura dai carichi di lavoro sul cluster locale o un problema di integrità del collegamento sta influenzando la latenza di uno o più carichi di lavoro sul cluster locale. Se il cluster non si trova in una configurazione MetroCluster, questa icona non viene visualizzata.

- **Operazioni aggregate o aggregate SSD**

  Rappresenta l'aggregato di storage su cui vengono eseguiti i carichi di lavoro. Se il componente aggregato è in conflitto, significa che un utilizzo elevato dell'aggregato influisce sulla latenza di uno o più carichi di lavoro. Un aggregato è costituito da tutti gli HDD o da una combinazione di HDD e SSD (un aggregato di Flash Pool). Un "Saggregato SD" è costituito da tutti gli SSD (un aggregato all-flash) o da una combinazione di SSD e un Tier cloud (un aggregato FabricPool).

- **Latenza cloud**

  Rappresenta il componente software del cluster coinvolto nell'elaborazione i/o tra il cluster e il livello cloud in cui vengono memorizzati i dati dell'utente. Se il componente di latenza del cloud è in conflitto, significa che una grande quantità di letture da volumi ospitati sul Tier cloud influisce sulla latenza di uno o più carichi di lavoro.

- **Sync SnapMirror**

  Rappresenta il componente software del cluster coinvolto nella replica dei dati utente dal volume primario al volume secondario in una relazione sincrona di SnapMirror. Se il componente Sync SnapMirror è in conflitto, significa che l'attività delle operazioni di SnapMirror Synchronous influisce sulla latenza di uno o più carichi di lavoro.

Ruoli dei carichi di lavoro coinvolti in un evento di performance

Unified Manager utilizza i ruoli per identificare il coinvolgimento di un workload in un evento di performance. I ruoli includono vittime, tori e squali. Un carico di lavoro definito dall'utente può essere una vittima, un bullo e uno squalo allo stesso tempo.

Ruolo	Descrizione
Vittima	Un carico di lavoro definito dall'utente le cui performance sono diminuite a causa di altri carichi di lavoro, detti "bulli", che utilizzano in modo eccessivo un componente del cluster. Solo i workload definiti dall'utente sono identificati come vittime. Unified Manager identifica i carichi di lavoro delle vittime in base alla loro deviazione nella latenza, in cui la latenza effettiva, durante un evento, è notevolmente aumentata rispetto alle previsioni di latenza (intervallo previsto).

Ruolo	Descrizione
Bully	Un workload definito dall'utente o dal sistema il cui utilizzo eccessivo di un componente del cluster ha causato la diminuzione delle performance di altri workload, denominati vittime. Unified Manager identifica i carichi di lavoro ingombranti in base alla loro deviazione nell'utilizzo di un componente del cluster, in cui l'utilizzo effettivo, durante un evento, è notevolmente aumentato rispetto all'intervallo di utilizzo previsto.
Squalo	Un carico di lavoro definito dall'utente con il massimo utilizzo di un componente del cluster rispetto a tutti i carichi di lavoro coinvolti in un evento. Unified Manager identifica i carichi di lavoro di Shark in base all'utilizzo di un componente del cluster durante un evento.

I carichi di lavoro su un cluster possono condividere molti dei componenti del cluster, come gli aggregati e la CPU per l'elaborazione di rete e dati. Quando un carico di lavoro, ad esempio un volume, aumenta l'utilizzo di un componente del cluster al punto che il componente non riesce a soddisfare in modo efficiente le richieste di carico di lavoro, il componente è in conflitto. Il carico di lavoro che sta utilizzando in eccesso un componente del cluster è un'operazione molto importante. Gli altri carichi di lavoro che condividono tali componenti e le cui performance sono influenzate dal problema sono le vittime. Anche le attività dei carichi di lavoro definiti dal sistema, come la deduplica o le copie Snapshot, possono essere sottoposte a escalation in "bullismo".

Quando Unified Manager rileva un evento, identifica tutti i carichi di lavoro e i componenti del cluster coinvolti, inclusi i carichi di lavoro ingombranti che hanno causato l'evento, il componente del cluster in conflitto e i carichi di lavoro vittime le cui performance sono diminuite a causa dell'aumento dell'attività dei carichi di lavoro ingombranti.

Se Unified Manager non riesce a identificare i carichi di lavoro ingombrante, avvisa solo sui carichi di lavoro vittime e sul componente del cluster interessato.

Unified Manager è in grado di identificare i carichi di lavoro vittime di carichi di lavoro ingombranti e di identificare anche i casi in cui questi stessi carichi di lavoro diventano carichi di lavoro ingombranti. Un carico di lavoro può essere un'attività molto ingombrante per se stesso. Ad esempio, un carico di lavoro dalle performance elevate che viene rallentato da un limite di gruppo di policy causa la limitazione di tutti i workload del gruppo di policy, anche se stesso. Un carico di lavoro ingombrante o vittima di un evento di performance in corso potrebbe cambiare il proprio ruolo o non essere più un partecipante all'evento.

Informazioni sul copyright

Informazioni sul marchio commerciale