Sommario del documento di prova

Sommario

Introduzione

Questo documento descrive la metodologia generale per risolvere i problemi relativi a un'interfaccia GUI APIC lenta.

Avvio rapido

Spesso si riscontra che i problemi lenti dell'interfaccia grafica APIC sono il risultato di un'alta percentuale di richieste API originate da uno script, un'integrazione o un'applicazione. Il file access.log di un file APIC registra ogni richiesta API elaborata. Il file access.log di un APIC può essere analizzato rapidamente con lo script Access Log Analyzer all'interno del progetto aci-tac-scripts del gruppo Datacenter di Github.

Premesse

APIC come server Web - NGINX

NGINX è il DME responsabile degli endpoint API disponibili su ciascun APIC. Se NGINX non è attivo, le richieste API non possono essere gestite. Se NGINX è congestionato, l'API è congestionata. Ogni APIC esegue il proprio processo NGINX, quindi è possibile che solo un

singolo APIC possa avere problemi con NGINX se solo tale APIC è oggetto di query aggressive.

L'interfaccia utente APIC esegue più richieste API per popolare ogni pagina. Analogamente, tutti i comandi show APIC (NXOS Style CLI) sono wrapper per script Python che eseguono più richieste API, gestiscono la risposta, quindi la forniscono all'utente.

Registri rilevanti

Nome file di log	Posizione	In quale supporto tecnico	Commenti
file access.log	/var/log/dme/log	APIC 3 di 3	ACI agnostic, fornisce 1 linea per richiesta API
errore.log	/var/log/dme/log	APIC 3 di 3	ACI Agnostic, mostra errori Inginx (limitazione inclusa)
nginx.bin.log	/var/log/dme/log	APIC 3 di 3	specifico di ACI, registra le transazioni DME
nginx.bin.warnplus.log	/var/log/dme/log	APIC 3 di 3	Contiene registri con livello di avviso + gravità

Metodologia

Isolamento trigger iniziale

Quali sono le conseguenze?

- quali APIC sono interessati; uno, molti o tutti gli APIC?
- Dove si vede la lentezza; tramite interfaccia utente, comandi CLI o entrambi?
- Quali pagine o comandi specifici dell'interfaccia utente sono lenti?

Come si sperimenta la lentezza?

- Questa condizione viene rilevata in più browser per un singolo utente?
- Più utenti segnalano la lentezza o solo un singolo utente o un sottoinsieme di utenti?
- Gli utenti interessati condividono una posizione geografica o un percorso di rete simile dal browser all'APIC?

Quando è stata notata per la prima volta la lentezza?

- È stata aggiunta di recente un'integrazione o uno script ACI?

- L'estensione del browser è stata abilitata di recente?
- Si è verificata una modifica recente nella configurazione ACI?

Verifica stato e utilizzo NGINX

Formato voce Access.log

access.log è una funzionalità di NGINX ed è pertanto indipendente da APIC. Ogni riga rappresenta una richiesta HTTP ricevuta da APIC. Fare riferimento a questo registro per informazioni sull'utilizzo di NGINX di un APIC.

Formato predefinito di access.log in ACI versione 5.2+:

```
log_format proxy_ip '$remote_addr ($http_x_real_ip) - $remote_user [$time_local]'
                    '"$request" $status $body_bytes_sent '
                    '"$http_referer" "$http_user_agent"';
```

Questa riga rappresenta una voce access.log quando viene eseguita una moquery -c fvTenant:

```
127.0.0.1 (-) - - [07/Apr/2022:20:10:59 +0000]"GET /api/class/fvTenant.xml HTTP/1.1" 200 15863 "-" "Pyt
```

Mappa della voce access.log di esempio a log_format:

campo log_format	Contenuto da esempio	Commenti
$remote_addr	127.0.0.1	IP dell'host che ha inviato la richiesta
$http_x_real_ip	-	IP dell'ultimo richiedente se proxy in uso
$utente_remoto	-	Generalmente non utilizzato. Selezionare nginx.bin.log per tenere traccia dell'utente che ha eseguito l'accesso per eseguire le richieste
$ora_locale	07/Apr/2022:20:10:59 +0000	Quando la richiesta è stata elaborata
$request	SCARICA /api/class/fvTenant.xml HTTP/1.1	Metodo Http (GET, POST, DELETE) e URI

$status	200	Codice di stato risposta HTTP
$body_bytes_sent	1586	dimensioni payload risposta
$http_referer	-	-
$http_user_agent	Python-urllib	Tipo di client che ha inviato la richiesta

Comportamenti di Access.log

Frequenza elevata di picchi di richieste in un lungo periodo di tempo:

- I burst continui di oltre 40 richieste al secondo possono causare rallentamenti dell'interfaccia utente
- Identificare gli host responsabili delle query
- Ridurre o disabilitare l'origine delle query per verificare se questo migliora i tempi di risposta di APIC.

Risposte 4xx o 5xx coerenti:

- Identificare il messaggio di errore da nginx.bin.log

Il file access.log di un APIC può essere analizzato rapidamente con lo script Access Log Analyzer all'interno del progetto aci-tac-scripts del gruppo Datacenter di Github.

Verifica utilizzo risorse NGINX

Per controllare l'utilizzo della CPU e della memoria NGINX, usare il comando top dell'APIC:

<#root>

```
top - 13:19:47 up 29 days, 2:08, 11 users, load average: 12.24, 11.79, 12.72
Tasks: 785 total, 1 running, 383 sleeping, 0 stopped, 0 zombie
%Cpu(s): 3.5 us, 2.0 sy, 0.0 ni, 94.2 id, 0.1 wa, 0.0 hi, 0.1 si, 0.0 st
KiB Mem : 13141363+total, 50360320 free, 31109680 used, 49943636 buff/cache
KiB Swap: 0 total, 0 free, 0 used. 98279904 avail Mem

PID   USER PR NI VIRT    RES   SHR   S %CPU %MEM TIME+     COMMAND
21495 root 20 0  4393916 3.5g 217624 S
```

2.6

2.8 759:05.78

nginx.bin

L'elevato utilizzo delle risorse NGINX può essere direttamente correlato a un elevato numero di

richieste elaborate.

Verifica core

Un arresto anomalo di NGINX non è tipico per i problemi relativi all'interfaccia grafica dell'APIC lento. Tuttavia, se vengono individuati core NGINX, collegarli a un TAC SR per l'analisi. Per la procedura di verifica dei core, consultare la guida ACI Techsupport.

Verifica latenza da client a server

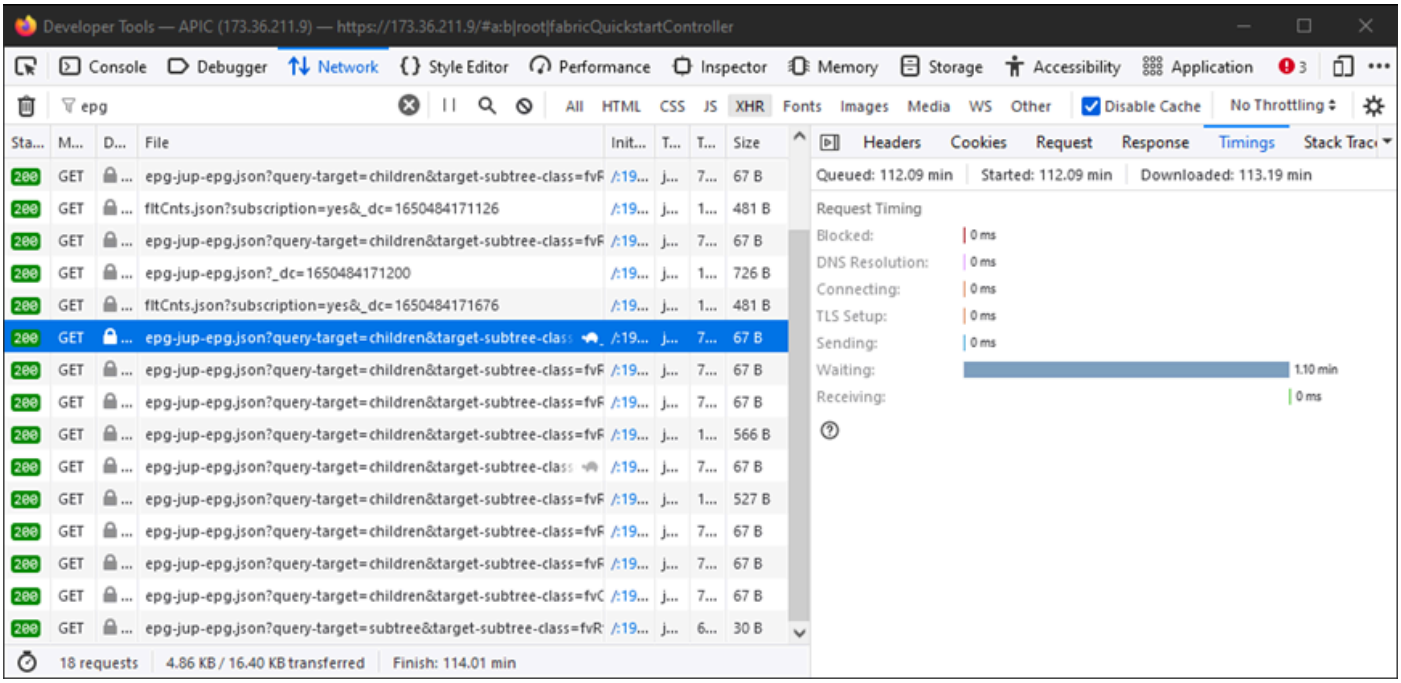
Se non vengono trovate richieste rapide ma un utente continua a mostrare lentezza dell'interfaccia utente, il problema può essere la latenza da client (browser) a server (APIC).

In questi scenari, convalidare il percorso dati dal browser all'APIC (distanza geografica, VPN, ecc.). Se possibile, distribuire e verificare l'accesso da un server di collegamento situato nella stessa area geografica o nello stesso centro dati degli APIC da isolare. Convalida se altri utenti mostrano una latenza simile.

Scheda Rete degli strumenti di sviluppo del browser

Tutti i browser sono in grado di convalidare le richieste e le risposte HTTP tramite il relativo toolkit di sviluppo dei browser, in genere all'interno di una scheda Rete.

Questo strumento può essere utilizzato per convalidare la quantità di tempo necessaria per ogni fase delle richieste originate dal browser, come mostrato nell'immagine.



Esempio di browser in attesa di una risposta da parte dell'APIC di 1,1 minuti

Miglioramenti per pagine specifiche dell'interfaccia utente

Pagina Gruppo di criteri:

L'ID bug Cisco CSCvx14621 - l'interfaccia grafica APIC viene caricata lentamente sui criteri IPG nella scheda Fabric.

Interfaccia nella pagina Inventory:

Cisco ID bug CSCvx90048 - Il carico iniziale della scheda operativa "Layer 1 Physical Interface Configuration" è lungo e induce al 'blocco'.

Suggerimenti generali per Client > Latenza server

Alcuni browser, come Firefox, consentono per impostazione predefinita più connessioni Web per host.

- Verificare se l'impostazione è configurabile nella versione del browser in uso
- Questo aspetto è più importante per le pagine con più query, ad esempio la pagina Gruppo di criteri

La velocità VPN e la distanza da APIC aumentano la lentezza complessiva dell'interfaccia utente in base alle richieste del browser client e al tempo di spostamento della risposta APIC. Un jump box geograficamente locale rispetto agli APIC riduce in modo significativo i tempi di spostamento del browser rispetto agli APIC.

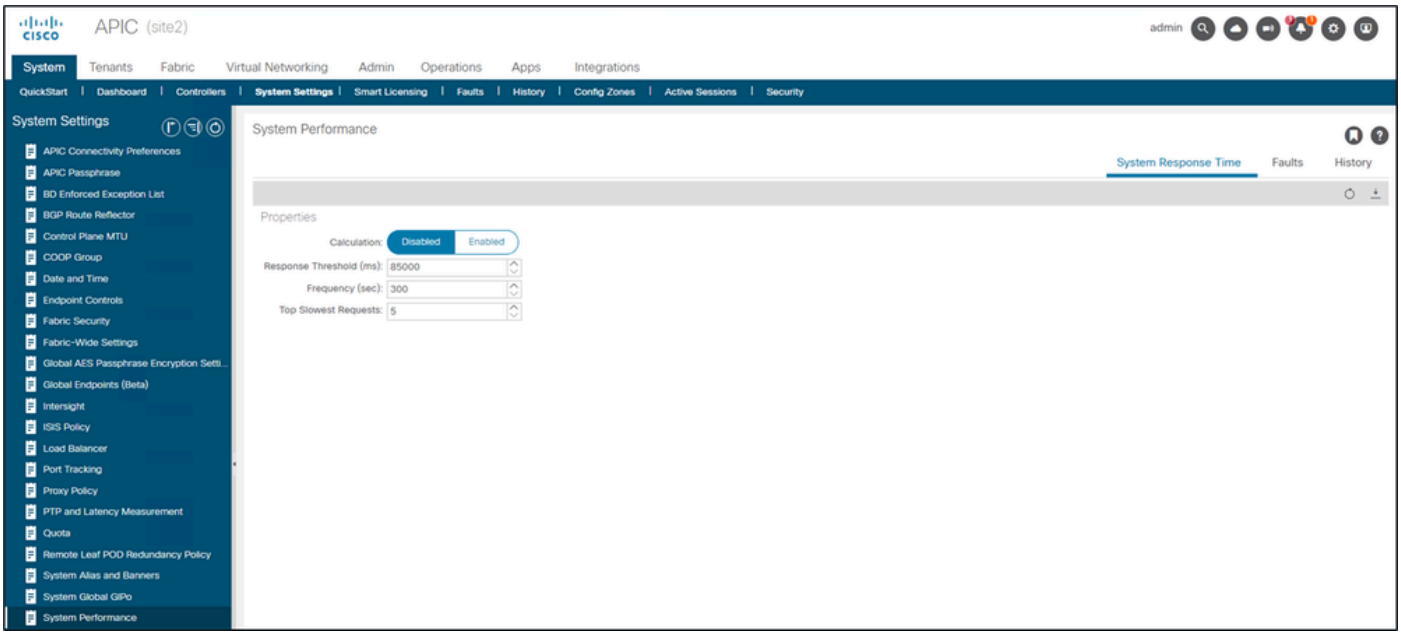
Verifica richieste Web lunghe

Se un server Web (NGINX su APIC) gestisce un volume elevato di richieste Web lunghe, ciò può influire sulle prestazioni di altre richieste ricevute in parallelo.

Ciò è particolarmente vero per i sistemi che dispongono di database distribuiti, ad esempio APIC. Una singola richiesta API può richiedere ulteriori richieste e ricerche inviate ad altri nodi nell'infrastruttura, con tempi di risposta più lunghi previsti. Una frammentazione di queste richieste Web lunghe in un intervallo di tempo ridotto può aumentare la quantità di risorse richieste e comportare tempi di risposta inaspettatamente più lunghi. Inoltre, le richieste ricevute possono scadere (90 secondi), determinando un comportamento imprevisto del sistema dal punto di vista dell'utente.

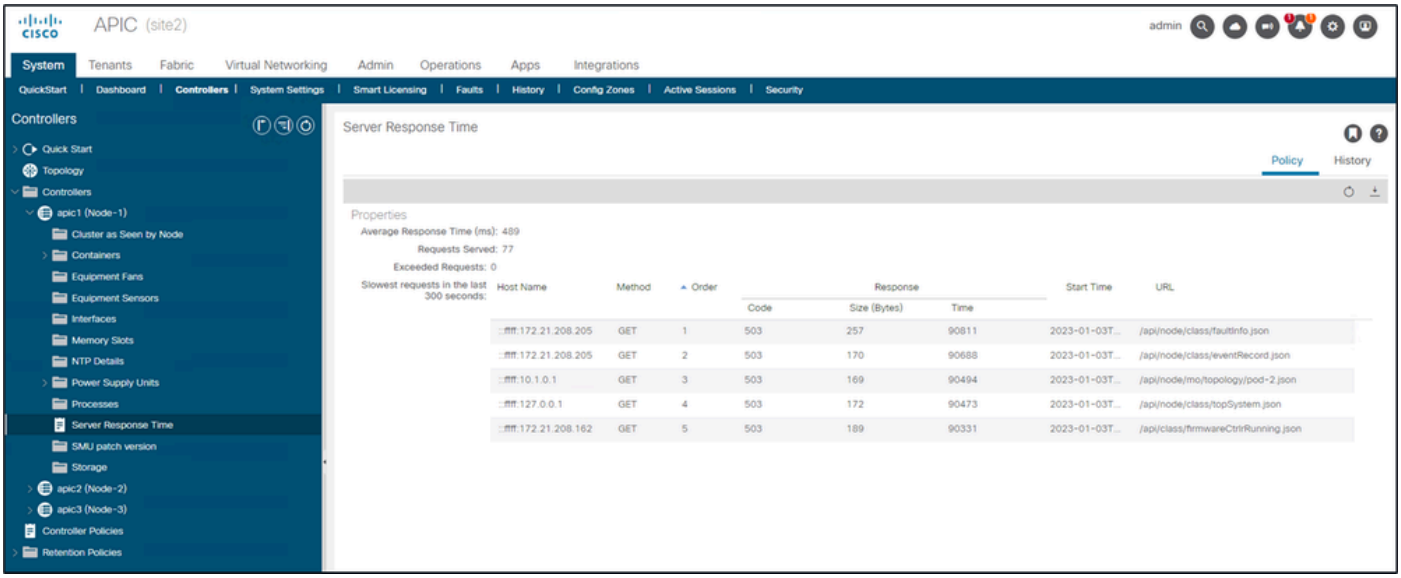
Tempo di risposta del sistema - Abilita calcolo per il tempo di risposta del server

In 4.2(1)+, un utente può abilitare il "Calcolo delle prestazioni del sistema" che tiene traccia ed evidenzia le richieste API che hanno impiegato molto tempo a gestire.



È possibile abilitare il calcolo da Sistema - Impostazioni di sistema - Prestazioni del sistema

Una volta abilitato il "calcolo", l'utente può passare a specifici APIC in Controller per visualizzare le richieste API più lente negli ultimi 300 secondi.



Sistema - Controller - Cartella controller - APIC x - Tempo di risposta server

Considerazioni sull'utilizzo delle API APIC

Puntatori generali per garantire che uno script non danneggi Nginx

- Ogni APIC esegue il proprio NGINX DME.
  - Solo l'NGINX di APIC 1 elabora le richieste all'APIC 1. L'NGINX di APIC 2 e 3 non elabora tali richieste.
- In generale, 40+ richieste API al secondo in un lungo periodo di tempo debilita NGINX.
  - Se individuato, ridurre l'aggressività delle richieste.
  - Se l'host Richieste non può essere modificato, prendere in considerazione i [limiti di

velocità NGINX sull'APIC.

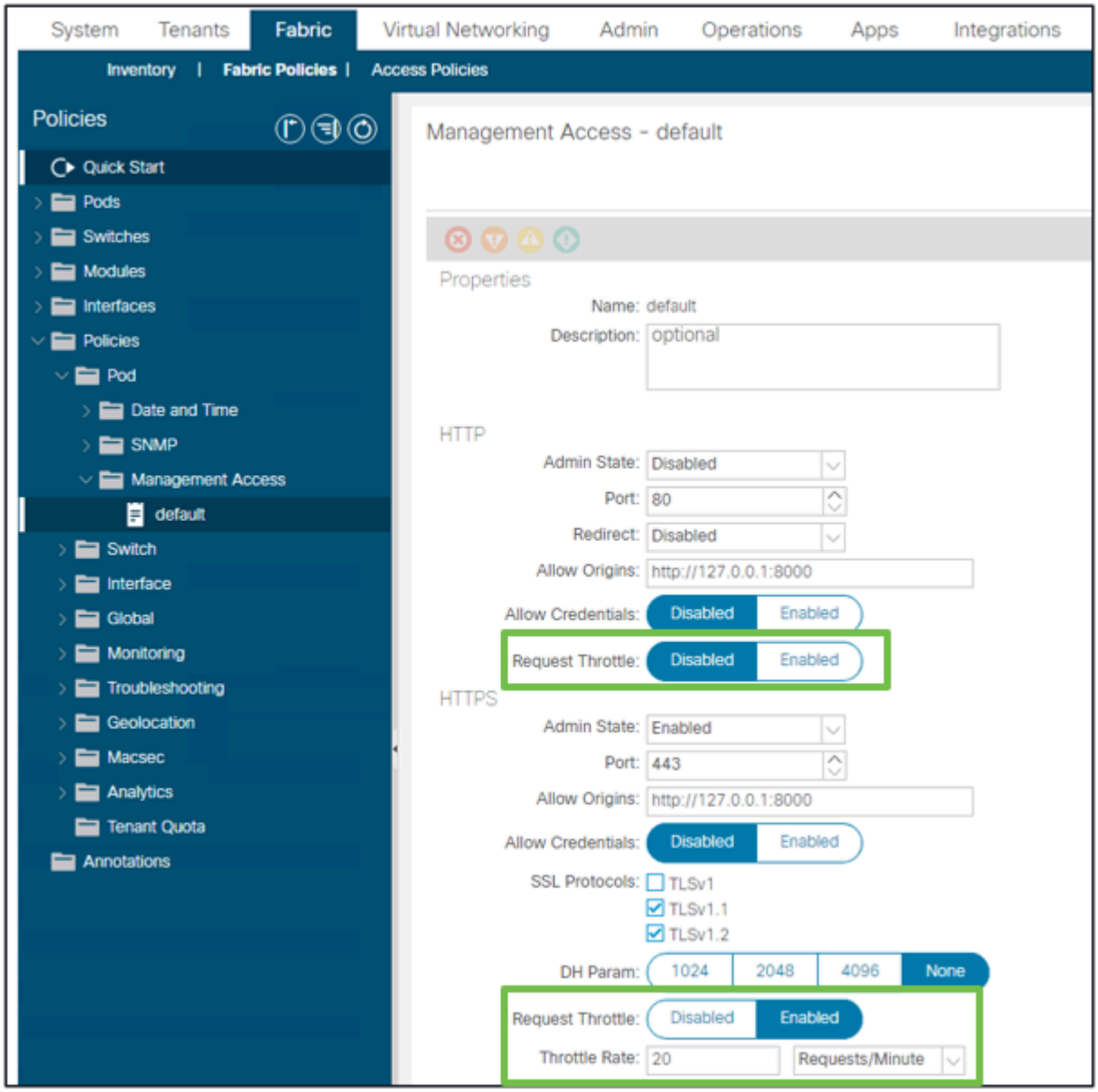
Inefficienze degli script di indirizzo

- Non eseguire l'accesso/la disconnessione prima di ogni richiesta API.
  - Il timeout predefinito per una sessione di accesso è 10 minuti. La stessa sessione può essere utilizzata per più richieste e può essere aggiornata per estendere il periodo di validità.
  - Vedere Guida alla configurazione dell'API REST Cisco APIC - Accesso all'API REST - Autenticazione e gestione di una sessione API.
- Se lo script esegue query su molti DN che condividono un padre, anziché comprimere le query in una singola query padre logica con Filtri query.
  - Vedere Guida alla configurazione dell'API REST Cisco APIC - Composizione delle query dell'API REST - Applicazione dei filtri di ambito delle query.
- Se è necessario aggiornare un oggetto o una classe di oggetti, prendere in considerazione le sottoscrizioni websocket anziché le richieste API rapide.

Limitazione richieste NGINX

Disponibile nella versione 4.2(1)+, un utente può abilitare la limitazione delle richieste su HTTP e HTTPS in modo indipendente.

Nota: a partire dalla versione ACI 6.1(2), la velocità massima supportata per questa funzionalità è stata ridotta a 40 richieste al secondo (r/s) o 2400 richieste al minuto (r/m) da 10.000 r/m.



Fabric - Criteri fabric - Cartella criteri - Cartella accesso gestione - predefinito

Quando abilitato:

- NGINX viene riavviato per applicare le modifiche al file di configurazione
  - Una nuova zona, httpsClientTagZone, viene scritta nella configurazione di Inginx
- La velocità può essere impostata in Richieste al minuto (r/m) o Richieste al secondo (r/s).
- La limitazione delle richieste si basa sull'implementazione del limite di velocità inclusa in NGINX
  - Le richieste API sull'interfaccia /api/URI utilizzano la velocità definita dall'utente + burst= (velocità x 2) + nodelay
    - Esiste una limitazione non configurabile (zona aaaApiHttps) per /api/aaaLogin e /api/aaaRefresh che limita la velocità a 2r/s + burst=4 + nodelay
  - La velocità delle richieste viene registrata in base all'indirizzo IP del client
  - Le richieste API originate dall'indirizzo IP automatico APIC (UI + CLI) ignorano la limitazione
  - Qualsiasi indirizzo IP del client che supera la velocità definita dall'utente + soglia di burst riceve una risposta 503 dall'APIC

- Questi 503 possono essere correlati nei log degli accessi
- error.log include voci che indicano quando è stata attivata la limitazione (zona httpsClientTagZone) e per quali host client

<#root>

apic#

less /var/log/dme/log/error.log

...
2023/04/17 20:19:14 [error] ...

limiting requests

, excess: 40.292 by zone "

httpsClientTagZone

", client: h.o.s.t, ... request: "GET /api/class/...", host: "a.p.i.c"
2023/04/17 20:19:14 [error] ...

limiting requests

, excess: 40.292 by zone "

httpsClientTagZone

", client: h.o.s.t, ... request: "GET /api/node/...", host: "a.p.i.c"

Come regola generale, Request Throttle serve solo a proteggere il server (APIC) da sintomi simili a quelli di DDOS indotti da client che eseguono query. Comprendere e isolare il client di richiesta per le soluzioni finali nella logica dell'app/script.

Consigli

Queste raccomandazioni sono progettate per contribuire a ridurre il carico e lo stress operativo sull'APIC, in particolare negli scenari in cui nessuna singola origine è responsabile di un elevato volume di chiamate API. Implementando queste procedure ottimali, è possibile ridurre al minimo l'elaborazione, la registrazione e la generazione di eventi non necessari nel fabric, migliorando la stabilità e le prestazioni del sistema. Questi suggerimenti sono particolarmente rilevanti in ambienti in cui i comportamenti aggregati piuttosto che gli incidenti isolati contribuiscono al ceppo APIC.

Disabilita registrazione ACL

Verificare che la registrazione ACL sia disattivata durante le normali operazioni. Abilitarlo solo durante le finestre di manutenzione pianificata per la risoluzione dei problemi o il debug. La registrazione continua può generare un numero eccessivo di messaggi informativi, in particolare con il traffico di volumi elevati su più switch, che aumenta il carico di lavoro APIC.

Per ulteriori informazioni, consultare la guida alla configurazione della sicurezza di Cisco APIC (collegamento della guida 5.2.x): https://www.cisco.com/c/en/us/td/docs/dcn/aci/apic/5x/security-configuration/cisco-apic-security-configuration-guide-release-52x/security-policies-52x.html

Limita conversione syslog a eventi critici

Configurare il sistema in modo che solo i messaggi syslog di gravità ALERT vengano convertiti in eventRecords. Evitare di convertire il livello INFORMATION (inclusa la registrazione ACL) per evitare che gli eventi rumorosi sovraccarichino l'APIC:

1. Passare a Fabric → Fabric Policies → Policies → Policies → Monitoring → Common Policy → Syslog Message Policies → Default.
2. Regolare il filtro della funzione per impostare la gravità del syslog su alert.

Codici degli eventi non essenziali Squelch

Eliminare (squelch) i codici di evento che non sono rilevanti per il monitoraggio per ridurre il rumore.
Per squelch del codice di evento E4204939, usare questo comando su qualsiasi CLI APIC:

```
bash
icurl -k -sX POST -d '<fabricInst><monCommonPol><eventSevAsnP code="E4204939" sev="squelched"/></monCom
```

Per verificare:

```
bash
icurl -k -sX GET 'https://localhost/api/node/class/eventSevAsnP.xml' | xmllint --format -
```

In alternativa, controllare tramite interfaccia utente:
Fabric > Criteri fabric > Criteri > Monitoraggio > Criteri comuni > Criteri di assegnazione della gravità degli eventi

Ottimizza aggiornamenti sottoscrizioni DNS

Per i fabric gestiti da versioni ND precedenti alla 3.2.2m o alla 4.1.1g, eseguire l'aggiornamento a una di queste versioni o successive per ottimizzare gli intervalli di aggiornamento delle sottoscrizioni. Le versioni precedenti vengono aggiornate ogni 45 secondi per unità MO, con una

scalabilità che può generare oltre 300.000 richieste APIC al giorno. Le versioni aggiornate aumentano il timeout di sottoscrizione a 3600 secondi (1 ora), riducendo gli aggiornamenti a circa 5.000 al giorno.

Monitoraggio delle query correlate a Intersight

I fabric abilitati a Intersight generano query periodiche sui topsistemi dal connettore DC (ogni 15 secondi), aggiungendo al carico APIC.
Nella release 6.1.2 e successive, questa query è stata ottimizzata per ridurre il sovraccarico.

Regola criteri di conservazione per i record

Impostare il criterio di conservazione per eventRecord, faultRecord e healthRecord su 1.000 per evitare un accumulo eccessivo di record. Ciò è particolarmente utile quando si estraggono questi record regolarmente per una specifica attività operativa. Valutare sempre l'impatto della riduzione della granularità del monitoraggio rispetto ai requisiti operativi e di risoluzione dei problemi.

Informazioni su questa traduzione

Cisco ha tradotto questo documento utilizzando una combinazione di tecnologie automatiche e umane per offrire ai nostri utenti in tutto il mondo contenuti di supporto nella propria lingua. Si noti che anche la migliore traduzione automatica non sarà mai accurata come quella fornita da un traduttore professionista. Cisco Systems, Inc. non si assume alcuna responsabilità per l’accuratezza di queste traduzioni e consiglia di consultare sempre il documento originale in inglese (disponibile al link fornito).