Flash talk: Wide Area Swift

Matteo Panella - matteo.panella@lngs.infn.it

INFN - Laboratori Nazionali del Gran Sasso

Miniworkshop CCR 2013

Sommario

Sommario

Chi, cosa e perché

Chi siamo

Un "sottoinsieme" del Cloud WG (LNGS, Bari, Padova, in futuro Roma2)

Cosa stiamo facendo

Un cluster OpenStack Swift distribuito su scala geografica

Perché

- trovare soluzioni tecniche per un repository centralizzato di immagini VM
- valutare l'utilizzo di Swift come object storage distribuito su scala geografica per uso generale

Per approfondire

Workshop gruppo Cloud INFN - M. Panella, “Casi d’uso e proposte tecniche per un repository nazionale di immagini VM in Glance”
`http://goo.gl/KVbBnX`

Sommario

Setup attuale

- 3 proxy node (LNGS, Bari, Padova)
- 4 storage node (1x100GB LNGS e Bari, 2x100GB Padova)
- 1 load balancer HAproxy (LNGS, un ulteriore LB pianificato a Bari)
- 1 server Keystone (LNGS)

Cosa funziona (bene)

- replica geografica trasparente dei dati (attualmente in replica 3)
- fault tolerance (condizioni anomale dello storage, problemi di rete...)
- fault recovery **automatico** da perdita totale dello storage su singoli nodi

Cosa funziona (meno bene)

- modesto impatto sulla rete (alto numero di connessioni tra storage node)
- impatto significativo sulle risorse CPU degli storage node (sono stati osservati load average ≈ 7 su una media di 15 minuti)
- documentazione per gli amministratori carente o contraddittoria
- ACL da configurare sui firewall di frontiera relativamente complesse

Cosa non funziona (come vorremmo)

- il port range dei servizi degli storage node si sovrappone a quello di X11 (!!!)
- i file di configurazione vanno mantenuti (grossomodo) identici **su tutto il cluster**...
- ...quindi è necessario implementare un sistema di configuration management
- Keystone è (attualmente) un single point of failure dell'intero cluster

Sommario

A che punto siamo

- Glance riesce ad accedere direttamente a Swift
- l'indicizzazione delle immagini va fatta **manualmente** dall'amministratore
- le installazioni CloudStack e OpenNebula sono "tagliate fuori" per mancanza di supporto diretto

Quindi?

C'è bisogno di un middleware che si occupi dell'indicizzazione automatica delle immagini e che sia compatibile con diverse infrastrutture cloud.

vmcaster e vmcatcher

vmcaster e vmcatcher sono due software open source sviluppati da HEPiX che permettono la creazione ed il consumo di “podcast” di immagini di VM:

- implementati in Python
- trasporto su HTTP(S)
- indipendenti dall'infrastruttura cloud utilizzata (plugin per OpenStack Glance sviluppato da EGI)
- licenza Apache License Version 2.0

vmcaster/vmcatcher e Swift

- vmcatcher può recuperare immagini e image list direttamente da Swift via HTTP
- vmcaster non supporta l'upload su Swift...
- ...ma stiamo sviluppando una patch per aggiungere il supporto

Modello ideale del repository

- upload diretto su Swift tramite vmcaster
- download ed aggiornamenti automatici con vmcatcher (eseguito via cron o simili)
- ???
- profit!

Sommario

Keystone

Keystone rappresenta il single point of failure dell'intero cluster: senza i servizi di autenticazione e autorizzazione di Keystone, Swift semplicemente smette di funzionare.
È necessario individuare soluzioni che permettano di scalare Keystone su più siti geografici:

- replica active/passive con DRBD su WAN
- replica active/active con HAproxy e Percona XtraDB/Galera su WAN

Keystone (2)

Presso LNGS è in test un mini-cluster Percona XtraDB distribuito su 2 siti geografici (LNGS e GSSI).
I test preliminari (load e check di alcuni milioni di righe) hanno dato buoni risultati, anche in caso di anomalie sul cluster.

Prossimo passo

Implementare un cluster Galera in replica 3 su LNGS, Bari e Padova e migrare Keystone su questo cluster.

Swift

Per Swift sono necessarie diverse cose:

- maggiori risorse di calcolo (hardware dedicato)
- fine-tuning dei parametri di replica e object auditing
- test con workload differenti (ad es. con copie di dati di esperimenti)
- implementare una “central authority” per la gestione della configurazione (in particolare i ring file)
- redigere documentazione accurata su configurazione e best practices per setup distribuiti su WAN

Attività del gruppo

Le attività prioritarie del gruppo di lavoro al momento sono la replica di Keystone e l'implementazione del supporto Swift in vmcaster.
Le prossime attività riguarderanno principalmente l'uso "generalista" di Swift su WAN come object storage distribuito e fault tolerant. Il contributo di altre sedi a questo tipo di sperimentazione è più che benvenuto ☺

Grazie per l'attenzione!