Deployment geografico di GPFS

Alessandro Brunengo
Mirko Corosu
INFN-Genova

Contenuto

- Mirror sincrono geografico
- Active File Management

Mirror sincrono geografico

- Soluzione per disaster recovery
- Configurazione di tre siti geograficamente distinti
  - Due siti contengono le risorse di nodi e storage per contenere una replica completa del file system
  - Il terzo sito consiste di un nodo ed un disco, utilizzati per garantire il quorum
  - Il cluster GPFS e' unico, distribuito sui tre siti
- In caso di failure di un sito
  - il quorum e' garantito dalla disponibilita' dei due rimanenti siti
  - i meccanismi di failover di GPFS e la disponibilita' di una replica completa garantiscono la funzionalita' del cluster
- La replica secondaria e' mantenuta in due possibili modi
  - data e metadata replica feature di GPFS
    - non richiede supporto di specifico hardware per lo storage
  - Mirror sincrono utlizzando IBM TotalStorage Enterprise Storage Server® (ESS) Peer to Peer Remote Copy

Repica sincrona di GPFS

- Singolo cluster, con nodi e dischi sui due siti
- Ogni disco di uno stesso sito ha stesso failure group
- I file system vengono configurati con replica factor 2 per dati e metadati
- I due siti che ospitano dati avranno uguale numero di quorum node
- Il terzo sito ha la funzione critica di mantenere il quorum per il cluster e per il file system descriptor
  - un singolo nodo: quorum node (garantisce il quorum in caso di failure di un sito)
  - un singolo disco interno, configurato come NSD, servito dal nodo del sito, con usage: descOnly
    - questo dice a GPFS di mettere su quel disco solo il file system descriptor, che garantisce il quorum per i descriptor del file system
  - il disco deve essere in un terzo failure group

Vantaggi e svantaggi

- Questa configurazione e' resiliente alla failure di uno qualunque dei tre siti
  - in modo trasparente, grazie alla gestione del failover di GPFS
- Questa soluzione limita la performance in funzione della latenza tra i due siti che ospitano repliche
  - questa penalizzazione puo' essere inaccettabile in funzione dei requisiti delle applicazioni
  - soluzione consigliata entro i 100 km

Active File Management

- AFM e' un meccanismo di file system caching integrato in GPFS
- AFM permette di creare una associazione tra un file system GPFS locale ed un file system remoto (GPFS o NFS), di definire il flusso dei dati e di disporre di un namespace globale
- AFM permette di mascherare la latenza geografica e connection failure
  - e' possibile continuare ad accedere e modificare dati anche in caso di failure
  - il caching consente di effettuare modifiche localmente, che vengono riportate in modo asincrono, eliminando la latenza di un accesso remoto alle applicazioni

AFM Home e Cache

- Home: file system remoto
  - file system esportato via NFS, che puo' essere un fileset GPFS o un filesystem qualunque (NFSv3)
  - filesystem GPFS acceduto nativamente (tramite remote-cluster access)
    - in questo caso i due cluster devono essere opportunamente configurati per consentire l'accesso in remote cluster
- Cache: un AFM-enabled fileset locale utilizzato per fare da cache ai dati della Home
  - il fileset e' servito da uno dei gateway node designati sul cluster locale
  - questo nodo si definisce MDS (metadata server) del fileset, ed e' l'owner del fileset
  - gli altri gateway node contribuiscono al flusso di dati da e verso la Home
- La Cache legge dati dalla Home in seguito a richieste di read di un file, o a seguito di un prefetch (automatico o manuale)
  - ogni volta che un file e' letto localmente, viene scaricato dalla Home e salvato sul fileset della Cache
  - AFM supporta ACl ed extended attributes (da Home GPFS)
- Possono essere configurate piu' Cache (su cluster differenti) della stessa Home
- Uno stesso cluster puo' avere un fileset Cache di una Home remota, ed un fileset Home montato come Cache altrove

Modalita' di caching

- Read only: i dati locali (Cache) sono readonly: non e' permesso scrivere sul fileset della Cache
- Local update: i dati locali (Cache) sono readonly, ma sono permesse creazione di file o modifiche di file locali
  - update locali non vengono mai propagati alla Home
  - i file creati o modificati vengono marcati come locali e non piu' confrontati con la copia della Home

Modalita' di caching (cont.)

- Single writer: una sola Cache e' modificabile, e le modifiche sono riportate nella Home
  - non c'e' enforcing: se un file della Home viene modificato, AFM rileva l'inconsistenza
    - il fileset della Cache passa da stato Active a NeedResync
    - il comando resync deve essere eseguito (automaticamente in alcuni casi) per risolvere le inconsistenze
- Indipendent writer: piu' Cache della stessa Home sono scrivibili indipendentemente
  - non c'e' locking: le modifiche devono essere fatte in modo coerente ed e' compito dell'utente farlo
  - se due Cache modificano lo stesso file contemporaneamente, il risultato degli update della Home e' indefinito

Caching e sincronizzazione

- Operazioni asincrone
  - write, chmod, chown, create, mkdir, remove, rmdir, rename, link, symlink, attribute updates
  - gli update effettuati operando sulla Cache sono asincroni
    - il gateway puo' ritardare la propagazione delle modifiche per un tempo configurabile
    - file e directory vengono riconfrontati con le copie della Home ad intervalli di tempo configurabili
- Operazioni sincrone
  - read di dati e di directory

AFS caching

- Un file viene scaricato completamente dalla Home nella Cache quando ne vengono letti alcuni blocchi
  - l'operazione viene fatta anche se l'applicazione non richiede tutto il file (prefetch)
  - i dati vengono resi disponibili alla applicaizone via via che vengono letti
- Anche i metadati vengono salvati nella Cache
- La Cache effettua gli update sulla Home
  - quando scade il synchronization lag
  - quando una operazione sincrona dipende da operazioni asincrone in coda
  - manualmente (mmafsctl flushPending)

Cache eviction

- Ablitata per default, ed operante in base alla quota del fileset Cache
- Puo' essere eseguta manualmente (mmafmctl evict)
- Configurando opportunamente soft e hard quota, si puo' gestire la situazione in cui la Home scrive dati piu' rapidamente di quanto l'operazione di evict li cancelli

Disconnessione

- In caso di inaccessibilita' della Home:
  - le operazioni sincrone vengono gestite localmente
    - se relative a dati/metadati non presenti in Cache, generano I/O error
  - le operazioni asincrone vengono riscontrate alle applicazioni dai nodi gateway, ma tenute pending fino alla riconnessione
- AFM esegue log su disco delle operazioni pending dei gateway
  - in caso di failure di un gateway, il failover opera in modo che un altro gateway prenda carico delle operazioni in corso
    - il failover in caso di gateway failure e' procedura che opera solo se la Home e' GPFS
  - oviamente, per AFM fileset read-only o local-update la perdita di un gateway e' priva di conseguenze

Disabilitazione di AFM e migrazione dati

- Un fileset AFM puo' essere convertito in un fileset ordinario
  - il fileset deve essere unlinked per la disabilitazione di AFM
- AFS permette quindi di migrare i dati da un file system esportato NFS, su un filesystem GPFS
  - la migrazione non trasporta parametri specifici del file system Home, come quota, snapshot, policy, definizione di fileset.
  - la migrazione viene realizzata generando una lista di file nella Home, ed utilizzando un prefetch di questi file nella Cache (RO o LU)
- AFM e' anche utilizzato per implementare soluzioni di DR