Test e risultati sull'uso di un file system GPFS condiviso su rete WAN

Antonio Budano (INFN – Sezione di Roma 3)

In collaborazione con:

Vladimir Sapunenko (INFN - CNAF)

Elisabetta Vilucchi (INFN –LNF)

Sommario

- Introduzione
- Caratteristiche di GPFS di IBM
- Descrizione dell'architettura
- Test e risultati
- Conclusioni

Introduzione

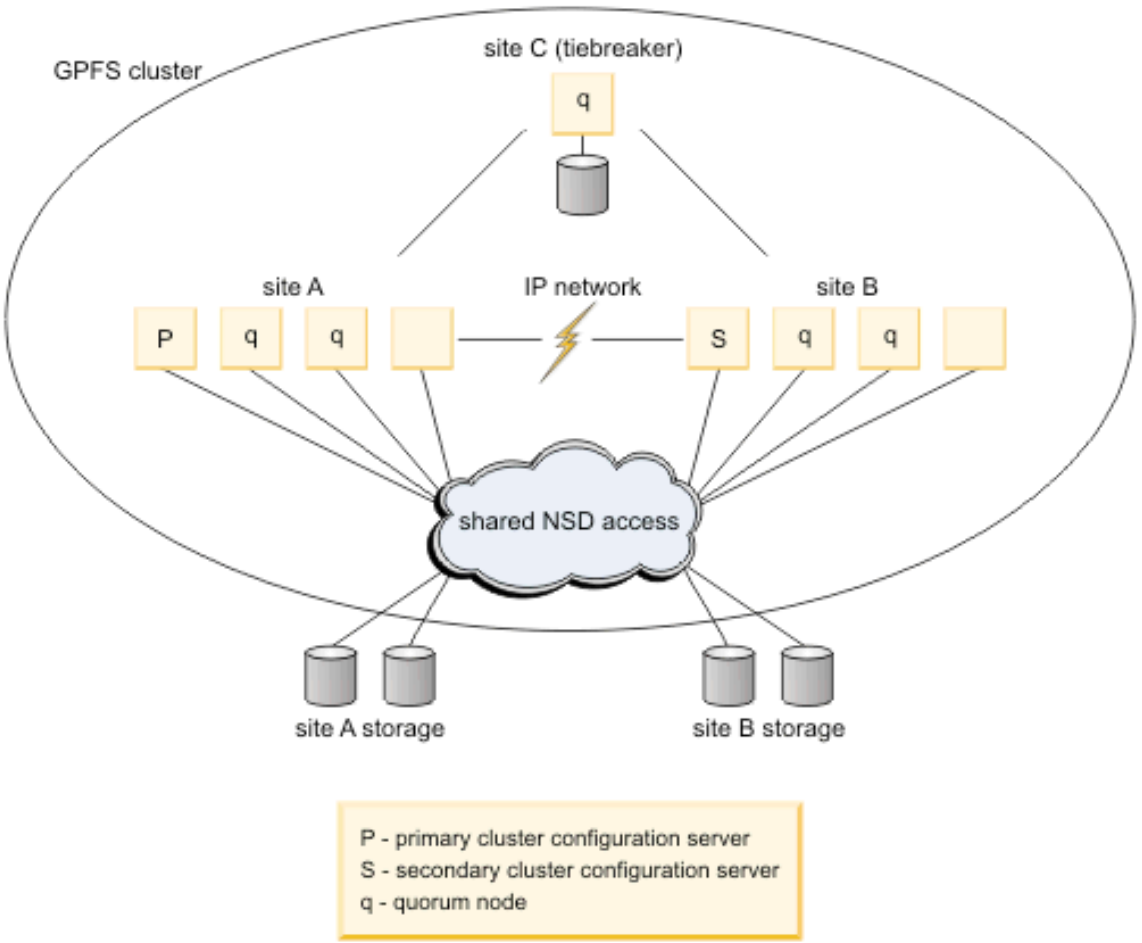
- L'idea di realizzare questa infrastruttura nasce nell'ambito del gruppo di lavoro dedicato a problematiche di Disaster Recovery.
- Necessità di poter avere un'area dove archiviare dati che sia affidabile e che contemporaneamente abbia delle performance sufficienti per le applicazioni che ne fanno uso come portali web, database e server virtuali.
- Abbiamo pensato ad un file system distribuito su rete WAN tra diverse sedi INFN.

Caratteristiche di GPFS (1/3)

- Perché GPFS?
  - GPFS (General Parallel File System) di IBM offre molte feature che permettono di facilitare l'implementazione di sistemi in alta affidabilità contro i guasti "catastrofici" sia hardware sia software
  - la possibilità di effettuare **repliche** dei dati in file system geograficamente separati
  - La possibilità di poter utilizzare funzioni di **Snapshot**
  - **AFM (Active File Management)** che permette di condividere i dati in maniera asincrona (adatto ad esempio per reti con alto valore di latenza) tra diversi cluster GPFS.

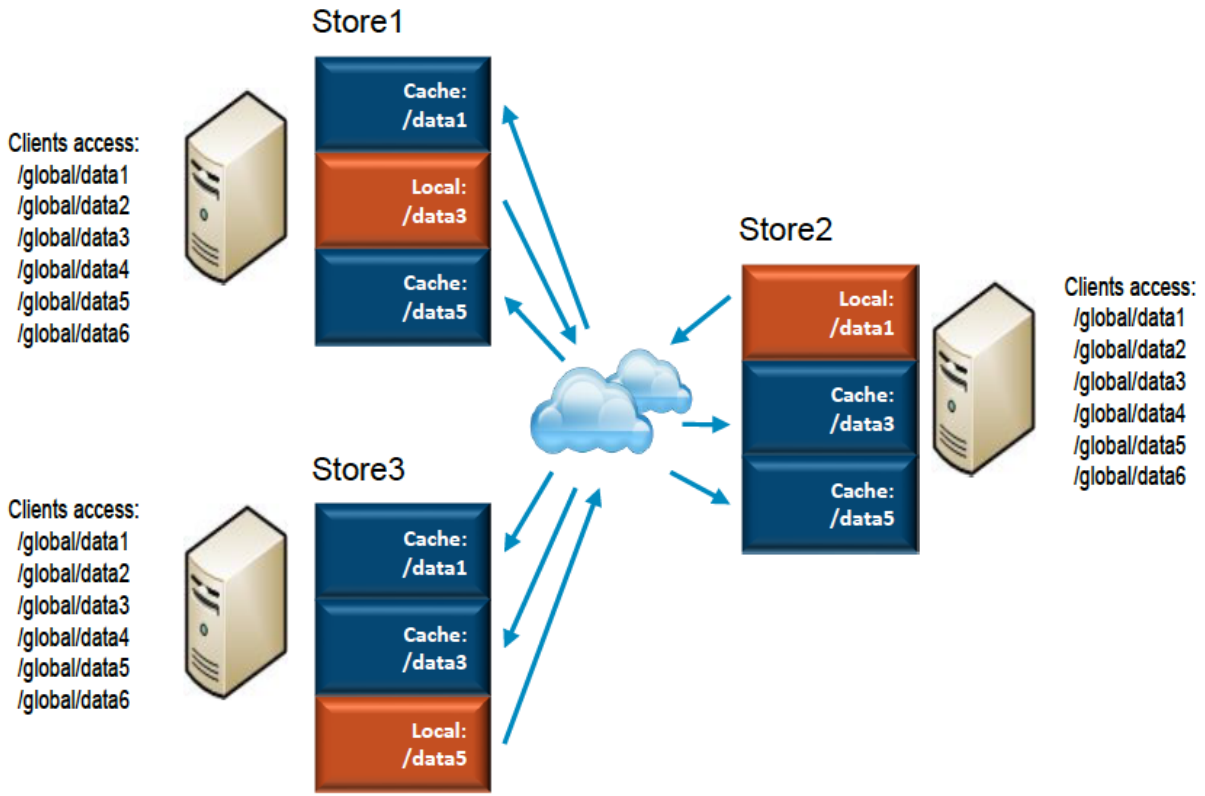
Caratteristiche di GPFS (2/3)

- GPFS permette di configurare il numero di copie di Dati e Metadati sui diversi storage presenti nel cluster. Questa caratteristica può essere usata per creare un synchronous mirroring tra coppie di siti geograficamente separati.

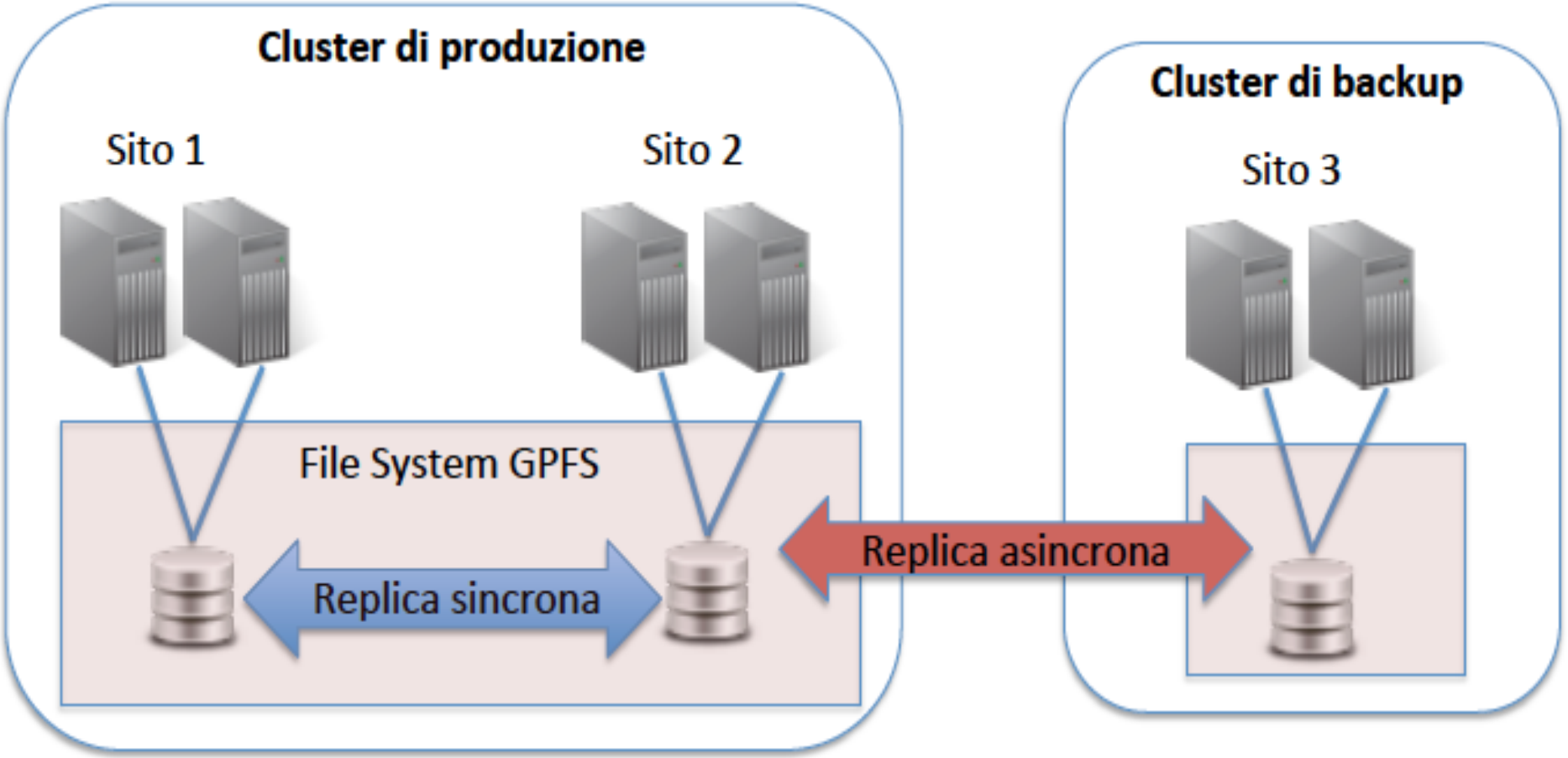


Caratteristiche di GPFS (3/3)

- Può essere utilizzata per creare global namespace tra diversi data center.
- AFM è pensato per permettere un efficiente trasferimento dati su rete WAN.
- Il trasferimento tra home -> cache può avvenire in parallelo trc i nodi



Architettura



Cluster di Produzione (1)

- CNAF – ROMA3
  - Bandwidth 1 Gbit/s, RTT=6ms
  - 2 servers + 2TB di storage in ogni sito
  - Scrittura Sincrona

Il parametro "read Replica Policy" permette di poter effettuare la lettura **direttamente sui dischi locali (se disponibili).**

- Test effettuati:
  - Scrittura e lettura sequenziale e Random utilizzando il pacchetto "Iozone"
  - Simulazione della rottura di uno dei due siti (disabilitando la porta ethernet nel server remoto) durante la scrittura di dati.
  - Simulazione della rottura del sistema di storage (disabilitando le porte Fiber Channel del sistema di storage) durante la scrittura di dati.
  - Utilizzando le applicazioni e i servizi del Sistema Informativo (portali web, database Oracle) residenti in Virtual Machine (KVM).

Rilsutati ed Osservazioni (1)

	Scrittura sequenziale 1MB blocks (MB/s)	Lettura sequenziale 1MB blocks (MB/s)	Scrittura Random 1MB blocks(MB/s)	Lettura Random 1MB blocks (MB/s)
Disco locale Roma 3	168	160	160	84
Disco locale CNAF	164	176	160	64
Replica del FS, da Roma3	51	168	74	55
Replica del FS dal CNAF	108	196	90	90

- Performance nella scrittura sequenziale sono limitate dalla bandwidth
- Lettura le performances sono limitate dal disco locale
- Scrittura e lettura Random limitazioni dovute alla latenza (RTT) tra i siti

Simulazioni di guasti

- Simulazione della rottura di uno dei due siti (disabilitando la porta ethernet nel server remoto) durante la scrittura di dati.
- Simulazione della rottura del sistema di storage (disabilitando le porte Fiber Channel del sistema di storage) durante la scrittura di dati.
- Osservazioni:
  - Fallimento di un sito durante una scrittura il tempo di ripristino dipende dalla configurazione del cluster che di default sono circa 60 s.
  - Tempo di recupero nel caso di fallimento del solo sistema di storage in uno dei due siti risulta pressoché immediato.

Cluster di Produzione (2)

- LNF – ROMA3:
  - Bandwidth tra Roma3 e Tier2 LNF: 10 Gbit/s, RTT=0,6ms (Rete privata progetto Megalab)
  - 2 servers + 2TB di storage in ogni sito
  - Scrittura Sincrona
- Test effettuati:
  - Scrittura e lettura sequenziale e Random utilizzando il pacchetto "Iozone"

Test e Risultati (2)

	Scrittura sequenziale 1MB blocks (MB/s)	Lettura sequenziale 1MB blocks (MB/s)	Scrittura Random 1MB blocks(MB/s)	Lettura Random 1MB blocks (MB/s)
Disco locale Roma 3	137	148	116	128
Disco locale LNF	118	390	118	384
Replica del FS, da Roma3	134	151	114	131
Replica del FS da LNF	118	382	115	378

- Le performances nella scrittura sul file system sono limitate dalla bandwidth tra i due siti.

Configurazione finale

- Aggiunto poi al cluster di produzione (2) un nuovo cluster al CNAF collegato via AFM (Cluster di Backup):
  - Home (ROMA3) – Cache (CNAF)
  - ROMA3 sito AFM Gateway
  - ROMA3 in questa configurazione è il solo sito di collegamento tra i due cluster (gateway).
- Test effettuati
  - Utilizzo delle applicazioni e dei servizi del Sistema Informativo (portali web, database Oracle) residenti in Virtual Machine (KVM)
  - Esecuzione di applicazioni sul cluster di produzione e cluster di backup

Conclusioni

- GPFS ci permette di realizzare una soluzione robusta di Disaster Recovery che contemporaneamente offre delle performance sufficienti alle nostre esigenze.
- La soluzione può garantire la continuità dei servizi: in caso di guasto ad uno dei due siti il sistema può essere ripristinato in pochi istanti
- Dal sito di backup (AFM) è possibile ripristinare il file system anche se entrambi i siti di produzione diventassero inaccessibili
- La possibilità di creare snapshot (ad esempio giornalieri) rende il sistema ancora più robusto:
  - è possibile, ad esempio, in caso di cancellazione accidentale di dati, ripristinare tutti i dati del file system a partire dagli snapshot precedenti.

Il Gruppo DR dell'INFN

Coordinatore : Stefano Zani

Componenti :
Sandro Angius
Massimo Donatelli
Claudio Galli
Guido Guizzunti
Dael Maselli
Massimo Pistoni
Claudio Soprano
Riccardo Veraldi
Vladimir Sapunenko
Stefano Bovina

+ Collaborazione di :
Nunzio Amanzi
Antonio Budano
Elisabetta Vilucchi
Alessandro De Salvo
(F.Sys. distribuiti)

Aree di intervento

DNS **{distribuito + geo-replica}**
MAILING **{distribuito + mail relay }**
SISTEMA INFORMATIVO :
- Contabilità (CNAF)
- Portale Utente (CNAF)
- Gestione Presenze (CNAF)
- Documentale (CNAF) [new]
- Business Intelligence BI (CNAF)
- Protocollo (CNAF) [new]
- AAI + GODIVA (LNF)
- Stipendiale Sxgest2 (LNF) [old]
- Stipendiale Cezanne (LNF) [new]
- Protocollo (LNF) [old]
- Documentale (LNF) [old]
- Portale Unico (LNF) [new]
- …

Ogni Target-IT INFN che sia considerabile «Core Business &/OR Mission Critical»