Gruppo storage CCR

Status Report

Alessandro Brunengo

Valutazione tecnologia iSCSI

- Cosa si voleva verificare
  - approfondimento delle specifiche e valutazione delle caratteristiche dell'hardware attualmente disponibile sul mercato
  - valutazione delle prestazioni di soluzioni con e senza HBA dedicate, impatto sulle risorse dei server
  - valutazione delle caratteristiche di fault tolerance dell'accesso al disco
  - test di prestazioni di controller iSCSI commerciali ed home-made
  - funzionalita' e scalabilita' su soluzioni di cluster file system (GPFS/Lustre)

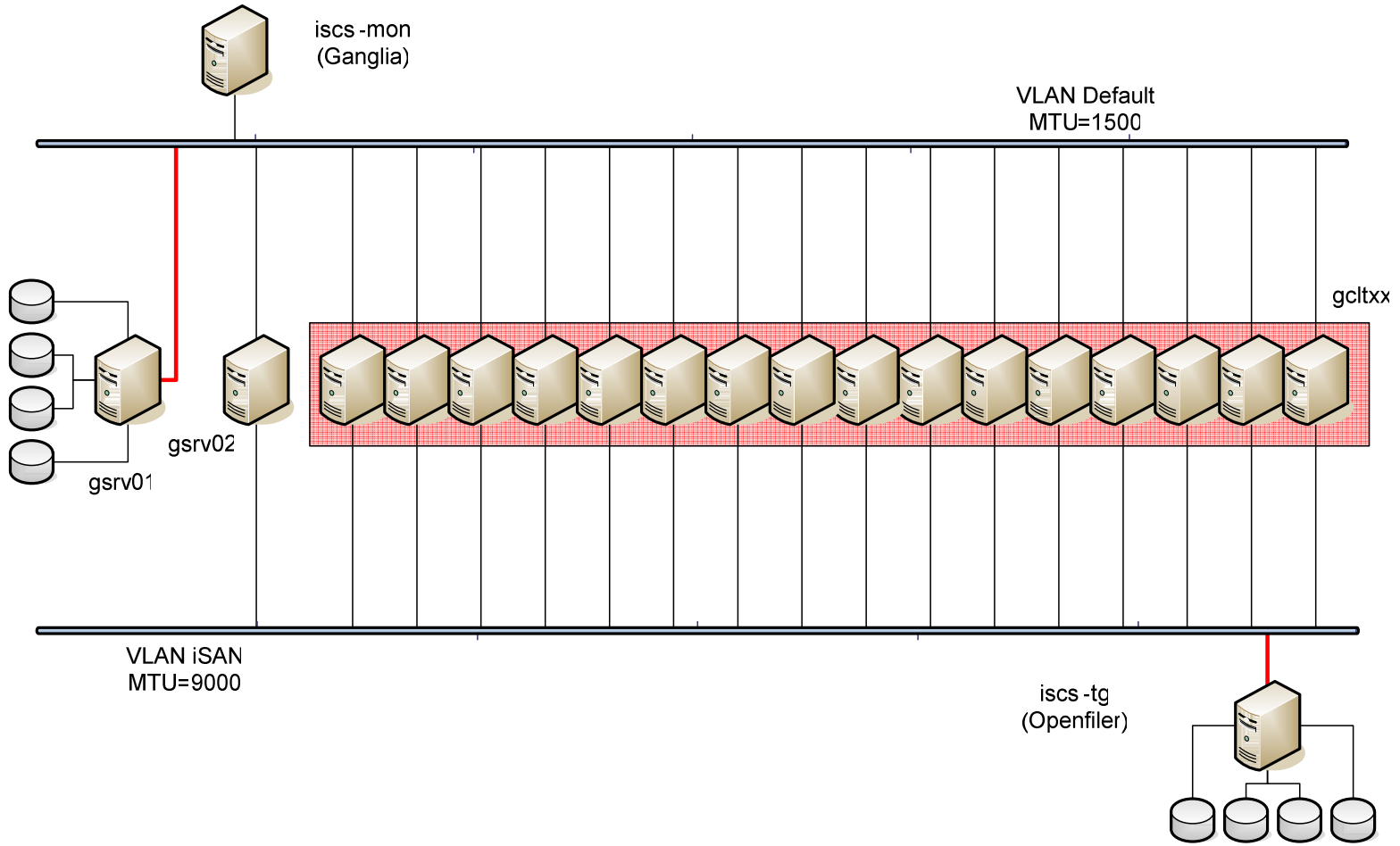
iSCSI: risultato della sperimentazione

- funzionalita' di accesso al disco e failover sul multipath: OK
- prestazioni con client/target senza HBA dedicata: NO
  - alto carico sulle CPU, con eventi di crash dell'OS
- oggetti commerciali
  - diffusi (quasi sempre con soluzioni in alternativa al FC), a costo equivalente
  - firmware di vari oggetti migliorato nel corso dell'anno
  - limiti sul numero di initiator e di sessioni contemporanee
  - prestazioni limitate al Gbps
- oggetti home-made (con HBA dedicata)
  - esistono soluzioni per realizzarlo (Openfiler)
  - prestazioni limitate al Gbps (equivalenti agli oggetti commerciali)
- restava da verificare:
  - link aggregation
  - utilizzo in ambiente GPFS per accesso diretto dei WN ai NSD

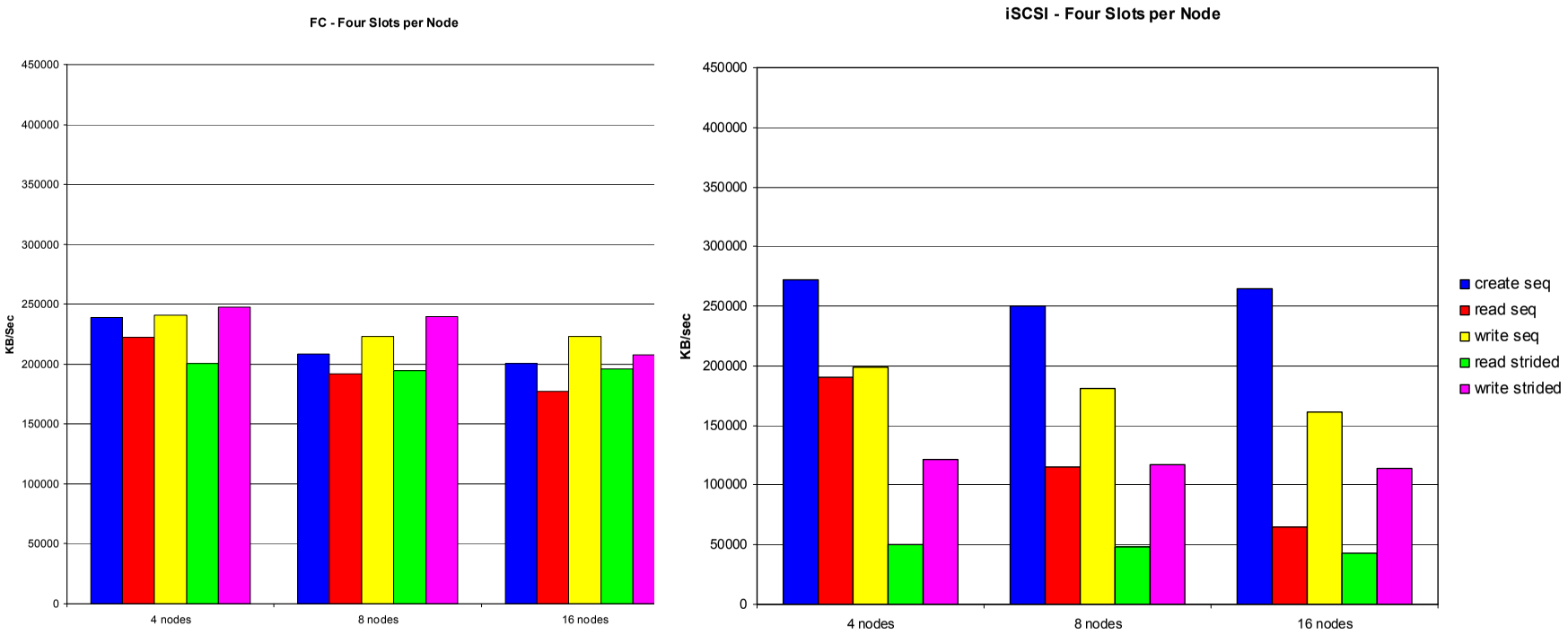
Test di iSCSI su link aggregation e file system parallelo (GPFS)

- Scelto algoritmo di selezione delle interfacce in funzione di MAC-src e MAC-dst
  - garantisce la sequenzialita' dei frame della singola comunicazione
- Eseguiti test di comparazione tra soluzione iSCSI+link aggregation e FC+link aggregation, su file system GPFS
- Misurate le prestazioni di traffico unidirezionale e bidirezionale

Testbed


iscs -mon
(Ganglia)
VLAN Default
MTU=1500
gcltxx
gsrv02
gsrv01
VLAN iSAN
MTU=9000
iscs -tg
(Openfiler)

Risultati comparativi



Considerazioni sulle ultime prove

- L'aggregazione diretta sul disk server presenta problemi
  - deve essere indagata: possibili problemi di TCP
  - non e' oggi una soluzione per superare il GE
- La funzionalita' di accesso diretto ai NSD di GPFS via iSCSI ha prestazioni peggiori di un accesso via NSD server
  - iSCSI non e' oggi una soluzione performante per l'accesso diretto agli NSD

Considerazioni conclusive sulla sperimentazione iSCSI

- iSCSI non sembra essere una soluzione SAN che possa sostituire FC dove siano richieste prestazioni elevate
- l'utilizzo senza HBA dedicata costituisce ulteriore fattore limitante
  - la soluzione non costituisce un risparmio rilevante
- Il suo utilizzo per l'accesso diretto ai volumi GPFS risulta penalizzante

- Puo' costituire una soluzione idonea in condizioni di prestazioni e carico contenuto (piccole installazioni)

- L'attivita' e' conclusa
  - un report definitivo sara' prodotto per la fine dell'anno
  - potrebbe riaprirsi in presenza di soluzioni commerciali a 10 Gbps

Interfacce SRM

- Novita' rispetto a ottobre
  - al Tier1 in produzione StoRM su file system GPFS da 100 TB (Atlas), ed in fase di test avanzato per LHCb
  - Altri siti Tier2 sono interessati a provare StoRM/GPFS (LNL)
  - Test effettuati al Tier1 con trasferimento via WAN (Cern) mostrano prestazioni a 270 MB/s sostenuti e 370 MB/s di picco su 4 teste GridFTP e 100 TB di file system

Attivita' in corso

- Valutazione del nuovo nameserver dCache (Chimera)
- Valutazione delle soluzioni di replica di file in dCache
  - la nuova release non la supportera' in un primo tempo, ma esistono soluzioni non ufficiali; l'ultima release reintrodurra' questa funzionalita' in febbraio
- Test di soluzioni di buffering per trasferimenti via WAN
- C'e' una proposta di costituire un centro di supporto nazionale su dCache in collaborazione con gli sviluppatori
  - supporto rapido
  - canale diretto con gli sviluppatori
  - interesse manifestato da Donvito/Spinoso, ma ancora da definire

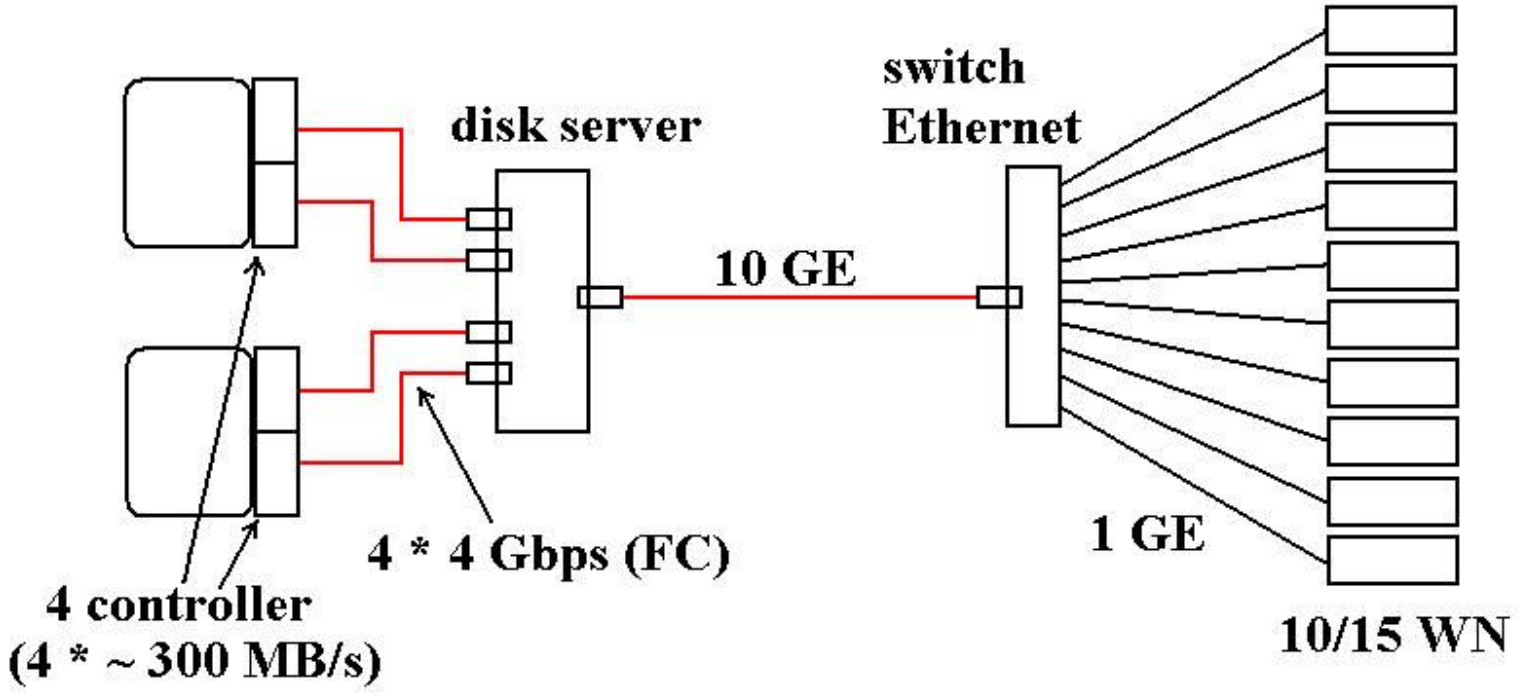
GPFS: attivita' in corso

- Utilizzo di policy per offrire robustezza al file system (in particolare in architetture prive di SAN)
  - repliche automatiche di metadati/dati
  - raggruppamento di disk pool per controllare la distribuzione dello striping dei file
  - eventuali policy di migrazione automatica di parti di dati in funzione di determinati eventi o caratteristiche
  - attivita' in corso di Vincenzo Vagnoni, ma altri siti interessati

Storage server a 10Gbps

- Proposta di nuova sperimentazione per il 2008
- Si vuole vedere quali prestazioni si possano ottenere da server connessi a 10 GE
  - capacita' del server di gestire un flusso di dati a 2 * 10 Gbps (da/verso disco e da/verso rete)
  - comportamento del TCP in condizioni di tale carico sul server
- Siti interessati: Genova, Pisa, Trieste, CNAF
- L'attivita' e' programmata in collaborazione con il netgroup

Layout del testbed



Risorse hardware

- Hardware disponibile
  - Due controller di disco a doppia testa con uscita in FC a 4 Gbps, 24 TB cad. con dischi SATA2 da 750 GB
    - opzione hardware in acquisto a Genova, disponibile fino a meta' marzo
    - alternativa oggetti in conto visione (Nexsan/Infortrend/...)
  - Due schede FC dual head a 4 Gbps PCI express (in acquisto a Genova e disponibli fino a meta' marzo)
  - 10/15 worker node con interfaccia in GE (Sez. di Pisa, periodo di disponibilita' da verificare)
- Hardware da trovare in conto visione
  - un server con almeno tre slot PCI express
  - una scheda 10 GE PCI express (acquisto?)
  - uno switch 24 porte GE ed uplink a 10 GE

Test e tempistica

- Test di prestazioni sulla parte di rete
  - netperf/iperf tra WN e server
- Test di prestazioni sul disco
  - dd a singolo flusso su ciascuno dei quattro canali FC server-disco (raw, ext3, GPFS)
  - dd con processi multipli server-disco (raw, EXT3, GPFS)
- Test di prestazioni di accesso remoto
  - dd a flusso multiplo WN-disco (GPFS)
  - applicativo di analisi (?? da identificare)
- Tempi previsti
  - periodo: febbraio 2008
  - durata: 5/10 giorni

Architetture di storage per Tier2

- Richiesta del managment:
  - produrre una panoramica delle possibili (ragionevoli?) architetture di storage per Tier2 con valutazione di robustezza, scalabilita', prestazioni, costi di acquisto e gestione
- Lavoro appena imbastito
  - brainstorm alla riunione del 10 dicembre
  - presenza di membri del gruppo di alcuni Tier2
    - non tutti

Matrice di possiblita'

- Molte opzioni su diversi aspetti
  - interfaccia SRM (dCache, DPM, StoRM)
  - accesso al file system (dCap, rfio, xrootd, file)
  - file system (locale: EXT3, XFS, remoto: NFS, o parallelo e distribuito: GPFS, Lustre)
  - infrastruttura hardware
    - HD (SATA, SAS/FC)
    - ridondanza (no RAID, RAID5, RAID6, controller con o senza failover)
    - accesso (DAS, SAN, SAN non switched)

      - Obiettivo primario: diminuire il numero di opzioni (3*4*5*2*3*2*3 = 2160)

- Non esiste una scelta giusta
  - eliminate le combinazioni che non possono funzionare, le altre possono (per definizione)

Su cosa basare la scelta?

- Esistono requisiti definiti
  - 500 TB, 500 slot, 60 MB/s in scrittura dal T1, 10 MB/s in lettura verso il T1, 1 GB/s in lettura verso i WN
- requisiti contrastanti
  - 5 MB/s/slot verso i WN fa 2.5 GB/s
- requisiti non ben definiti
  - quale livello di ridondanza (cioe' cosa e' accettabile perdere e per quanto tempo)?
- parametri la cui definizione e' soggettiva
  - utilizzare tecnologie note in un sito ha un costo di setup e gestione inferiore
- parametri la cui quantificazione e' complessa
  - facile valutare l'mtbf di un HD ed il costo indotto da una scelta
  - difficile valutare l'mtbf di un server con dischi ed il costo indotto
    - hardware migliore = mtbf maggiore (ma quanto?) = costo maggiore
  - costo: per la stessa soluzione tecnica e' grandemente dipendente dal vendor (migliore qualita', migliore supporto, minori tempi di disservizio)

Piano di lavoro

- Fare una grossa selezione in base alle caratteristiche delle parti
  - valutazione gia' avviata, da raffinare
- Identificare, per le diverse opzioni (che restano)
  - eventuali test da realizzare per verificarne funzionalita', prestazioni e scalabilita'
    - e raccogliere i dati di attivita' gia' svolte
  - costi (presunti) hardware in funzione del dimensionamento, che dipende dai requisiti (prestazioni, volumi)
  - vantaggi e svantaggi in termini di flessibilita', difficolta' di gestione, robustezza
    - difficilmente saranno numeri
- E' essenziale la collaborazione delle persone che operano sui Tier2, e potersi avvalere dell'esperienza del Tier1

Tempistica

- Il 18 dicembre avremo una prima stesura
  - sara' ovviamente incompleta
- Per meta' gennaio dovremo produrre un documento completo in termini di infrastrutture idonee e quantificazione di costi (ove possibile)
  - e proposte di test di scalabilita'

Corsi

- A fine novembre sono stati proposti due corsi di formazione
  - interfacce SRM: si ritiene importante realizzare un corso specifico dedicato a questo argomento, con approfondimenti tecnici e prove di installazione sul campo
    - Giacinto Donvito (per dCache) e Riccardo Zappi (per StoRM) si sono detti interessati a collaborare
    - sarebbe opportuno realizzare il corso in tempi non troppo lunghi (febbraio/marzo?): in corso di definizione
  - GPFS: la sua diffusione nell'INFN potrebbe beneficiare di un corso effettuato da uno specialista IBM (uno sviluppatore?) per affrontare in modo approfondito aspetti generali ma anche use case specifici
    - si deve ancora contattare IBM per valutare possibilita' e costi
    - tempistiche ovviamente da definire

Collaborazione con altri gruppi

- Prosegue la collaborazione con il gruppo storage di HepiX
  - G. Donvito userà i test che verranno effettuati sia a Bari sia con il resto del gruppo per fornire informazioni (principalmente su dCache) al gruppo dello storage di Hepix che sta creando un wiki con informazioni sulle varie tecnologie di storage (sia hardware che software)
- La collaborazione con il gruppo storage del Tier1 e' fondamentale
  - talvolta si riesce ad unire gli sforzi per realizzare test di scala e funzionalita' che richiedono necessariamente l'infrastruttura del T1
  - l'attivita' del gruppo storage T1 ha ricadute importanti sia per i Tier2 che per altri siti
  - c'e' qualche difficolta' per lo scambio delle informazioni, dovuto al sovraccarico di lavoro al T1 (si deve cercare di vedersi di piu')